IDENTIFYING ABERRANT RESPONDING: USE OF MULTIPLE MEASURES

A DISSERTATION SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Susan Christa Steinkamp

IN PARTIAL FULFILLMENT OF THE REQUIREUIREMENTS

FOR THE DEGREE OF

Doctor of philosophy

Michael C. Rodriguez, Adviser

May 2017

ACKNOWLEDGEMENTS

I would first like to thank my adviser, Dr. Michael Rodriguez, for his guidance and support throughout my doctoral studies. His feedback and advice were invaluable as well as his willingness to help me navigate through the challenges of being a non-traditional student.

I would also like to thank the members of my dissertation committee, Dr. Ernest Davenport, Jr., Dr. Chun Wang, and Dr. Andrew Zieffler, for their valuable insight and suggestions at the various stages of this study. Their feedback helped to improve the quality of my dissertation.

Special thanks goes to my friend and mentor, Dr. Kathleen Gialluca, for her never-ending encouragement and support. She inspired me to begin this journey and helped me through to the end.

I am grateful to Dr. Huijuan Meng for her technical expertise and generosity in sharing her time. I am fortunate to have such a good friend.

Finally, I would like to thank my parents, Chuck and Christa Stanek, who have always encouraged me to keep going and reach my goals. I am grateful for their love and support.

DEDICATION

To my sons, Sam and Jack, who studied with me and made me laugh. To my husband, Mike, whose patience, love, and support allowed me to complete my PhD. I couldn't have done this without him.

ABSTRACT

For test scores that rely on the accurate estimation of ability via an IRT model, their use and interpretation is dependent upon the assumption that the IRT model fits the data. Examinees who do not put forth full effort in answering test questions, have prior knowledge of test content, or do not approach a test with the intent of answering questions to the best of their ability are exhibiting aberrant response behaviors and the accuracy and validity of the resulting test scores are called into question. The test administrator is left with the problem of determining whether test scores are a true representation of examinee ability (Reise, 1990; Karabatsos, 2003). Model fit is typically assessed through item-fit indices. An equally important aspect of assessing model fit is determining how well an IRT model fits the response patterns of examinees, which is commonly referred to as person fit (Meijer & Sijtsma, 2001).

The purpose of this research was to explore the application of person-fit analysis in the identification of cheating behavior. Specifically, issues that may impact the effectiveness of person-fit indices, also called person-fit measures, were evaluated. A primary focus of this research was the value of using multiple types of measures (scalar, response time, graphical), both individually and combined, in determining whether or not a response pattern is indicative of cheating behavior.

A review of the literature on person-fit research is presented, followed by a discussion of considerations for designing a person-fit simulation study. A study was then conducted to determine the effectiveness of three person-fit measures in identifying simulated cheating behavior under various conditions. The person-fit measures used in the study were l_z (Drasgow, Levine & Williams, 1985), Effective Response Time (Meijer & Sotaridona, 2006), and the Person Response Curve (Trabin & Weiss, 1983). The effectiveness of the individual measures and the measures used in combination was evaluated. Study factors included IRT model, exam length, examinee ability level,

amount of aberrance within an exam, and amount of aberrance within a sample or population. A real-parameter simulation study (Seo & Weiss, 2013) was conducted using Rasch and two-parameter logistic (2PL) IRT parameters estimated from a large dataset obtained from a language skills assessment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Cheating on exams continues to be a growing problem in the assessment industry. It is not limited to a particular market (education, clinical, certification and licensure, etc.), population (age, race, social status, etc.), or geographical region. In secondary and post-secondary educational settings, cheating occurs across students of all ages and on all types of exams, from end-of-course and summative to college entrance exams. Cheating occurs by adults on professional licensure and certification exams, as illustrated in an article by Zamost, Griffin, and Ansari (2012) on the widespread cheating of radiology doctors on certification exams. The article describes how radiology residents memorize content while taking the test and then compile and share the items with others who are preparing to take the test. As stated by a representative from the American Board of Radiologists, the organization that oversees the certification program, not only is cheating on the exam a legal and ethical violation, but puts the safety of the public at risk.

The nature of cheating behavior has changed over the years with the transition from paper-pencil to computer-based test administration. Computer-based test (CBT) administration has provided a means for increased security of exam content with features like randomizing the order in which items are presented to examinees. Additionally, multiple versions or forms of an exam can be randomly assigned so examinees get different sets of items in addition to the items being presented in a unique order. Implementing such features in CBT administration makes it difficult, if not impossible, for examinees to copy answers from others or cue other examinees in the room to the correct answers. While CBT administration has deterred certain types of cheating behavior, other behaviors have taken their place. One example is item harvesting, which is defined by the International Test Commission (2014) as examinees memorizing test content to record and disseminate at a later time. When content is

harvested by one or more examinees, lists containing partial item content, correct answers or entire items verbatim can be produced. As illustrated above with the radiologists, harvesters sell, or even just give away, these lists to future examinees.

As one means of combating the issue of cheating, test administrators have looked for statistical methods to help in identifying compromised exam content and test-takers who may have cheated on an exam. An article by DiSario, Olinsky, Quinn, and Schumacher (2014) described a court case in which two individuals were suspected of cheating on a 100-item multiple-choice professional certification exam. The prosecution claimed that the two examinees had prior knowledge of the exam content, which resulted in each answering 93% of the items correctly. In addition, of the seven items answered incorrectly, the same incorrect answers were selected by the examinees for five items. Both the defense and prosecution employed independent consultants to complete statistical analyses of the exam results. Both consultants ran simulation studies to determine the probability of the defendants' exam results. In the end, statistical arguments were used by both the prosecution and defense lawyers to prove or disprove, respectively, that the examinees had cheated on the exam. The authors (who were on the side of the defense) asserted that without definitive proof and in light of the findings of the statistical analyses, subjective judgment is needed when trying to decide if cheating has occurred.

There is a large body of research regarding statistical methods used to identify cheating on exams and specifically, analysis of item response patterns to determine whether responses conform to an expected response pattern. Item responses that deviate from the expected response pattern are referred to as aberrant and can be an indication that cheating has occurred. However, making a determination about whether an examinee in a real testing situation is exhibiting aberrant response behavior, has and continues to be a difficult task (Drasgow, Levine, & Zickar, 1996; Meijer, 2003; St-Onge,

Valois, Abdous, & Germain, 2011). In addition, deciding what to do with this information when found can put the test administrator in a difficult position.

Cheating behavior left unchecked can be detrimental to a testing program. For example, replacing exposed item content is costly, requiring numerous resources to produce high quality items and extended periods of time to field test and calibrate those items. There are also costs that are harder to quantify, such as granting someone a professional license or certification based on a questionable test score and the risk to public safety or admitting a student into a higher education institution at the expense of a more qualified student. Alternatively, labeling or even implying that someone has cheated on an exam could result in monetary loss (e.g., loss of wages), denial of opportunity (e.g., admittance to higher education institution) or even psychological damage to that individual. Unlike item-fit analysis, where items that don't conform to the expected model can simply be set aside for further review or even discarded, labeling an examinee's score as invalid due to misfit has great consequences.

The importance of evaluating item response patterns for model fit (or aberrance) is evident. However, because the consequences of decisions based on such analyses (person-fit analyses) can be costly to both the test administrator and the examinee, the more information there is to support these decisions, the better. The use of multiple indicators to gather evidence of aberrance, rather than relying on a single measure, is strongly recommended (Meijer, 2003). Researchers and test administrators have numerous person-fit indices available to them and should select those most appropriate for a given testing program. In a computer-based testing environment, item and exam timing information can be used as one input to the evaluation of a response vector. Multiple indices can provide different viewpoints (e.g., scalar, timing, graphical) and strengthen the argument that a given response vector indicates aberrant behavior and may not be representative of the examinee's true ability level.

This study evaluated the effectiveness of currently available person-fit indices, individually and in combination, in identifying cheating behavior on an exam. Cheating was defined as examinees having knowledge of exam items prior to taking the exam (referred to as item pre-knowledge). This study also explored the impact of factors, such as underlying measurement model and certain exam characteristics, on the effectiveness of the person-fit indices. The primary goal was to determine whether gathering evidence from multiple points of view can provide better information to either support or dispel an assertion of cheating.

CHAPTER 2: REVIEW OF THE LITERATURE

Aberrant responding, simply put, is when a person answers hard items on a test correctly and easy items incorrectly. It is usually the case that aberrant responding is detected when a low ability individual gets many hard items right or a high ability individual gets many easy items wrong, beyond what would be expected by chance. In Item Response Theory (IRT), aberrant responding is defined as an observed item-response pattern that deviates from the expected response pattern based on a specific IRT model (Levine & Drasgow, 1988; Meijer & Sijtsma, 2001; Reise, 1990; Reise & Widaman, 1999; van Krimpen-Stoop & Meijer, 1999). Causes of aberrant responding in a testing environment include behaviors such as cheating (including copying responses from another examinee, having prior knowledge of exam items, and collusion), carelessness, low motivation, random responding, and guessing (Belov, 2013; Tatsuoka, 1984). These factors can affect ability estimates, but are unrelated to ability. In any case, the individual's item response pattern does not reflect that of the usual examinee who tries his/her best on the test, pays attention during the entire test, and has little difficulty understanding and interpreting the test questions.

To identify whether aberrant responding is present, statistical methods or indices (i.e., person-fit measures) have historically been used to assess the fit of an item-response pattern to an IRT measurement model. Person-fit measures provide an indication of how consistent an examinee's responses aggregated across items are with a specified IRT model (Reise, 1990). Person-fit indices can be used to answer questions such as "is an examinee behaving in a manner consistent with the model?" or "is the estimated ability an appropriate measure or representation of an examinee's true ability?" (de Ayala, 2009). Just as item fit is important to determine whether a given IRT model fits the data, person fit is important to determine whether an examinee is

responding to items according to the underlying construct being measured or whether there are other factors influencing response behavior (Meijer, 2003).

The focus of IRT model fit is typically on how well the model fits the data across the entire sample of examinees. With person fit, the focus is on how well the model fits at the individual or person level. To better contextualize person fit, Table 1 compares the reasons most often discussed in the research for assessing person fit to those for item fit.

Table 1. *Reasons for Assessing Item Fit and Person Fit*

Item fit	Person fit
Selecting an IRT model that best preserves the integrity of the observed data	Determining whether examinee response patterns are consistent with a specified IRT model
Confirming the unidimensionality of the data	Identifying and possibly removing misfitting response patterns to create a more unidimensional data matrix
Assessing item performance	Identifying examinees that are not measured well by a particular instrument
Identifying errors that may have occurred during calibration	

(Karabatsos, 2003; Reise, 1990; Reise & Widaman, 1999)

As can be seen in the table above, person fit (or misfit) can be useful not only in assessing the appropriateness of the IRT model, it can also address the appropriateness of the measurement of ability on which decisions are based (Karabatsos, 2003). Examinees with aberrant or misfitting response patterns are at risk of receiving inaccurate test scores which may lead to unfair or inappropriate decisions based on those test scores (de la Torre & Deng, 2008). In addition, having aberrant response vectors in the data matrix may have a negative influence on the calibration of items,

potentially affecting the ability estimates of a larger number of examinees. Recent research asserts that aberrant response patterns should be removed from the data in order to preserve the integrity of resulting item calibrations and validity of test scores (Belov, 2013).

Person-Fit Indices: Review of the Literature

There is a plethora of research on person-fit indices. Karabatsos (2003) and Meijer and Sijtsma (2001) provided comprehensive summaries of over 36 parametric and non-parametric appropriateness or person-fit indices. Most studies have focused on the theoretical and mathematical development of person-fit statistics and their power to detect aberrant response patterns (Meijer, 2003).

IRT-based indices are derived using the principles of IRT, such as the one-, two-, or three-parameter logistic models, the assumptions of local independence of items, uni-dimensionality of the measured trait, and typically, maximum likelihood parameter estimation. It is important to note that these IRT-based indices are computed with respect to a person's ability level. As Harnisch (1983) summarizes, "these indices indicate the extent to which examinees of equal ability differ in their pattern of responses" (p. 194).

Likelihood-based Indices

Likelihood-based indices measure the probability of an examinee's response pattern against the response pattern predicted by the model (Karabatsos, 2003; Reise & Widaman, 1999). One of the most widely used and researched likelihood-based statistics is l_z (Drasgow, Levine & Williams, 1985), which is the standardized version of l_o (Levine & Drasgow, 1982; Levine & Rubin, 1979). The l_o statistic is the log-likelihood of the response vector for an individual with ability estimate $\hat{\theta}$. Drasgow et al. (1985) provided the following equation for the computation of l_o:

$$l_o = \sum_{i=1}^{n}[u_i \log P_i(\hat{\theta}) + (1 - u_i) \log Q_i(\hat{\theta})] , \quad (1)$$

where

u_i is the item response (0 = incorrect, 1=correct) for item i (i = 1 to n),

$Q_i(\hat{\theta}) = 1 - P_i(\hat{\theta})$,

$P_i(\hat{\theta}) = \hat{c}_i + \frac{1+\hat{c}_i}{1+\exp[-D\hat{a}_i(\theta-\hat{b}_i)]}$,

D = 1.702 and $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ are item parameter estimates.

Because l_o is not a standardized statistic, the underlying null distribution is unknown. In addition, determining whether a response pattern is model-fitting (or misfitting) depends on the examinee's ability estimate. The l_z statistic was developed to overcome these issues and provide a measure of the likelihood of an examinee's response pattern relative to the expected response pattern at a given ability level (Reise & Widaman, 1999). The computation of the standardized statistic l_z is

$$l_z = \frac{l_0 - E(l_0)}{\sqrt{Var(l_0)}} \quad , \qquad (2)$$

where

$E(l_o)$ is the expected value of l_o and Var(l_o) is the variance of l_o (de Ayala, 2009, p. 143).

When the data fit the model, values of l_z should be near zero. Negative values of l_z are an indication of an inconsistent response pattern and positive values indicate that the response pattern is more consistent than the model predicts (Reise, 1990; de Ayala, 2009).

Molenaar and Hoijtink (1990) proposed the person-fit statistic M, which is a variation of l_o based on the Rasch model, as an alternative to l_z. Molenaar and Hoijtink (1990) argued that using an estimated theta rather than the true theta affects the distribution of the person-fit statistic (assumed to be normally distributed with a mean of 0 and a standard deviation of 1). Specifically, the variance of l_z can be smaller than expected, reducing its effectiveness in identifying aberrant response patterns. This is

not an issue under the Rasch model because the sum of the item scores (i.e., total raw score) is a sufficient estimator of ability.

Molenaar and Hoijtink (1990) show that l_o can be simplified as the sum of two terms as follows:

$$l_o = d_o + M, \quad (3)$$

where

$$d_o = -\sum_{i=1}^{n} \ln[1 + \exp(\theta - \mathrm{b_i})] + r\theta \ ,$$

and

$$M = -\sum_{i=1}^{n} \mathrm{b_i} X_i .$$

Given

$r = \sum_{i=1}^{n} X_i$, where X_i is the binary response for item i.

The term d_o is a constant across all item response vectors at a given value of r. In addition, d_o is independent of the item response vector (X_i is absent from the computation of d_o), whereas M is dependent on it. Because r, l_o and M have the same ordering in the item response vector (as given by l_o and M) and due to its computational simplicity, Molenaar and Hoijtink (1990) proposed using M rather than l_o as a person-fit statistic. They also provide various methods for approximating the distribution of M, emphasizing the utility of the chi-square distribution in which the mean, standard deviation and skewness of M are taken into account (Molenaar & Hoijtink, 1990; Meijer & Sijtsma, 2001).

Residual-based Indices

Like the mean-squared residual-based indices used to assess model-data fit at the item level, such indices have been developed to assess model-data fit at the person level. Wright (1977), an advocate of the Rasch model, proposed evaluating test scores for the effects of behaviors such as guessing, carelessness, or random responding by computing how much is “left over” after the data have been used to estimate item

difficulties and person abilities. How much is "left over", or the residual, is computed by taking the difference of the probability of a correct response from the actual item response (i.e., 1 for a correct response and 0 for an incorrect response). The standardized sum of squared residuals form an approximate chi-square distribution and can be used to test the appropriateness or aberrance of an individual's response pattern.

Two commonly-cited indices developed by Wright and Stone (1979) and Wright and Masters (1982) are the *U* and *W* statistics, respectively. The *U* statistic is computed for an individual by taking the average of the squared difference between the observed and expected responses over all items, divided by the conditional variances of the item scores. This statistic can be interpreted as the mean squared standardized residuals given *n* items (Karabatsos, 2003; Meijer & Sijtsma, 2001). Meijer and Sijtsma (2001) provide the following equation for the computation of *U*:

$$U = \sum_{i=1}^{n} \frac{[X_i - P_i(\theta)]^2}{nP_i(\theta)[1 - P_i(\theta)]} \quad (4)$$

The *W* statistic is also computed as the average of the squared item residuals, but weighted by the sum of the item variances (Karabatsos, 2003). Meijer & Sijtsma (2001) noted that the *W* statistic was assumed to be less influenced by unexpected responses to items with locations farther away from an individual's ability estimate. The following equation is provided for the computation of *W* (Meijer & Sijtsma, 2001):

$$W = \frac{\sum_{i=1}^{n}[X_i - P_i(\theta)]^2}{\sum_{i=1}^{n} P_i(\theta)[1 - P_i(\theta)]} \quad (5)$$

To evaluate *U* and *W* against a standard normal distribution, Wright and Stone (1979) and Wright and Masters (1982) transformed the statistics using a *Z*-cubic root (*Z*) or logarithmic transformation (*ln*). Transforming *U* and *W* using *Z* or *ln* yields the following statistics (Karabatsos, 2003): *ZU, ZW, lnU, lnW*.

Optimal Person-Fit Statistics

There are two person-fit statistics that are referred to as *optimal* person-fit statistics because they provide the most powerful tests of the null hypothesis that an item response pattern fits a specified model (versus the alternative hypothesis of misfit). The first statistic, $\lambda(\mathbf{X})$ proposed by Levine and Drasgow (1988), is a likelihood ratio statistic. The second statistic, $T(\mathbf{x})$ proposed by Klauer (1991, 1995), uses the exponential family of models to model specific types of aberrant response patterns and depends on which type of aberrance is being modeled (Meijer & Sijtsma, 2001). Although these measures of person-fit provide a statistically optimal method for identifying aberrant response patterns, they are difficult to implement because a model for both fit and a particular type of misfit must be specified. As noted by Drasgow, Levine, and Zickar (1996), to correctly utilize an optimal person-fit statistic, a quantitative model must be developed for a given set of items on a test and the model must be developed to reflect the "unique characteristics of the particular type of aberrance" (p. 53) under investigation. While there are significant benefits to using optimal person-fit statistics, the implementation and maintenance of these statistics in practice require a great deal of time and resources, making them less optimal for use in operational testing programs. More recently, the most common use of this type of statistic cited in the literature has been for testing dimensionality and local independence (Karabatsos, 2003; Meijer, 2003; Meijer & Sijtsma, 2001).

Person Response Curve

Trabin and Weiss' "person response curve" (PRC) and how it can be used to detect aberrant responding, is a method that makes most use of the characteristics found in the normal ogive curve and of the information that can be derived from the curve.

Trabin and Weiss (1983) described how to derive the PRC using the principles of IRT. Their method of detecting unusual response patterns used an observed PRC and an expected PRC. The observed PRC is obtained by plotting the proportion of items answered correctly as a function of item difficulty, for an individual with a given ability level. However, because the observed PRC is affected by the interaction of the individual with the items, it alone cannot tell you whether factors other than ability are affecting the individual's responses (i.e., whether the individual's test-taking behavior deviates from usual test-taking behavior).

The expected PRC is derived using the same IRT model (i.e., the one-, two-, or three-parameter model) and the same methods as those used to derive item characteristic curves in IRT. As Trabin and Weiss (1983) described, items are first ordered by difficulty and divided into groups (called strata) according to difficulty. The expected PRC is then constructed from the estimated probability of a correct response for each item and averaging the probabilities within strata. This allows the comparison of the observed PRC to the expected PRC to determine if the individual's responses fit the IRT model.

Trabin and Weiss (1983) explained how characteristics found in the shape of the observed PRC and that the degree to which it deviates from the expected PRC can indicate aberrant responding. These explanations come directly from basic assumptions and principles in IRT. The shape of the observed PRC can potentially provide information about a person's test performance like carelessness and guessing, or dimensionality. Carelessness may be present when a person is answering very easy items incorrectly (i.e., according to his/her ability level, he/she should be answering these questions correctly). Alternatively, a person correctly answering harder items on the exam (i.e., items that are above his/her ability level) may be a sign of guessing behavior. Another explanation for an examinee answering items in a manner not

consistent with his/her ability level is that the exam is not measuring a unidimensional construct as intended. That is, some other trait may be influencing test-taking behavior.

In addition to inspection of the PRC to identify aberrant response patterns, Trabin and Weiss (1983) developed a chi-square goodness-of-fit statistic. The chi-square statistic is computed using the expected and observed average probabilities of correct responses for each strata. The number of strata (minus 1) provides the degrees of freedom for assessing significance of the resulting chi-square statistic.

Extended Caution Indices

The Extended Caution Indices (ECI), developed by Tatsuoka and Linn (1983), are a family of caution indices based on Sato's caution index, which uses sample summary statistics and the observed response pattern for detecting unusual response patterns (Tatsuoka, 1984). The ECIs utilize IRT modeling and compare the item response pattern for an individual at a given θ level against the expected probability provided by the IRT model. The indices ECI4 and ECI5 are specifically categorized as individual caution indices and measure the degree to which an examinee's observed item response pattern relates to the theoretical person response curve (PRC) at a given θ level. These indices are similar to Trabin and Weiss' (1983) method of using chi-square to determine the relationship between item response patterns and a PRC derived from an IRT model.

The ECI4 is obtained by computing the covariance between the response vector $\mathbf{X}_i$ and $\mathbf{P}(\theta)$ as follows (Meijer & Sijtsma, 2001):

$$\text{ECI4} = 1 - \frac{Cov[X_i, P(\theta)]}{Cov[G, P(\theta)]}, \quad (6)$$

where

$\mathbf{X}_i$ is the item response vector for examinee *i*,

$\mathbf{P}(\theta)$ is the vector of the conditional probability of a correct response given θ,

G is (G1, G2, G3,…, G_k) with elements $G_g = {}^{1}/_{n} \sum_{i=1}^{n} P_g(\theta)$.

Finally, as with the l_o statistic described above, the ECI4 is confounded by theta level which can produce inflated statistical values at the high and low ends of the theta continuum. To account for this, Tatsuoka (1984) created a standardized version, denoted ECI4z, by subtracting the expected values of ECI4 and dividing by the standard deviation.

Detection of Aberrant Responding

While there is no general consensus on which method is best for assessing person fit, there is agreement in the literature regarding the causes of aberrant response patterns. The primary behavioral causes include cheating (answer copying, having prior knowledge of exam items, collusion), carelessness, low motivation, random responding, and guessing (Belov, 2013; Tatsuoka, 1984). Other possible causes of aberrant response patterns that have been investigated include local item dependence and differences due to demographic characteristics such as ethnicity or gender (Brown & Villarreal, 2007; Karabatsos, 2003). Although many studies have shown the utility of person-fit indices in detecting aberrant response patterns, identifying which cause to attribute aberrant response patterns to continues to be a challenge.

In addition, there are numerous factors that can impact the performance, and therefore the use, of person-fit indices. Factors such as test length, item parameters (spread of item difficulty and/or discrimination), IRT model used to estimate item and person parameters, and amount of aberrance exhibited in a response vector (St-Onge et al., 2011) have all been noted in the literature. As discussed previously, the type of aberrance (e.g., cheating, random responding) can also have an impact on detection rates, with some indices requiring that aberrance type be specified in advance.

Several studies have found that person-fit indices tend to perform better with longer tests and tests with items of varying difficulty levels (McLeod & Lewis, 1999; Molenaar & Hoijtink, 1990; Reise & Due, 1991). A longer test provides more opportunity for detecting deviant response patterns, assuming the items are reliable indicators of the construct of interest (Molenaar & Hoijtink, 1996). Similarly, a range of item difficulty levels is important because misfit is more easily detected when examinees are responding aberrantly to items with difficulty levels farther away from their ability level.

The empirical distribution of a person-fit index can also have an impact on its effectiveness in detecting aberrant responding. Classification of response patterns relies on an assumed theoretical distribution of a given person-fit index (typically assumed to be a standard normal distribution). If the empirical distribution does not approximate the theoretical distribution, the ability of a person-fit index to accurately classify response patterns as aberrant or non-aberrant will be diminished.

Case in point is the likelihood-based statistic l_z (Drasgow, Levine & Williams, 1985), which is one of the most widely researched and utilized person-fit statistics. While l_z has been the subject of a substantial amount of research and proven to be one of the most powerful statistics in detecting person misfit, the theoretical distribution of l_z is rarely achieved using real data (Nering, 1997; Seo & Weiss, 2013). Research has shown that factors such as range of item difficulty, item discrimination, test length, and the use of estimated θs versus true θs have an impact on the l_z distribution (Molenaar & Hoijtink, 1990; Nering, 1995; Seo & Weiss, 2013; van Krimpen-Stoop & Meijer, 1999).

When l_z is not distributed normally, hypothesis testing based on a standard normal distribution cannot be used to identify examinees with significantly misfitting response patterns. For example, van Krimpen-Stoop & Meijer (1999) noted that empirical distributions of l_z can differ from the assumed standard normal distribution for short tests, resulting in an empirical Type I error rate smaller than the nominal rate. In a

study by Seo and Weiss (2013), the authors used real data and estimated item parameters from an achievement test to evaluate the distribution of l_z and determine its accuracy in identifying person misfit. They found that item difficulty distribution and method used to estimate theta impacted the l_z distribution. Specifically, the authors suggested that a wide, rectangular distribution of item difficulty values and maximum-likelihood estimation of theta be used to determine the l_z distribution and critical values for detecting misfit. In addition, Seo and Weiss (2013) recommended "Monte Carlo simulation be implemented using the item parameters estimated with real data and an appropriately modeled distribution of θ" (p. 1014) to determine the observed l_z distribution and appropriate critical values for identifying aberrance.

St-Onge et al. (2011) explored the effect of various factors on the detection rates of parametric and non-parametric person-fit indices. Most notable, the researchers found that there was not a linear relationship between degree of aberrance in a response vector and the detection rate by a given index. Rather, there was a point at which increasing the degree of aberrance in a vector actually resulted in a decrease in detection rates. For example, when cheating behavior was simulated, it was found that detection rates for parametric indices (i.e., l_z) peaked when approximately 30% - 40% aberrance was simulated. The authors postulated that indices like l_z are highly dependent upon the ability estimate to detect aberrant responding so when a response vector has a high degree of aberrance, accurate ability estimation becomes difficult, impacting the effectiveness of the person-fit index.

When making a determination regarding the appropriate and most effective person-fit index to employ, it is important to not only explore the statistical characteristics of a person-fit measure, but also practical considerations. In an operational testing program, there are many aspects of the test development and delivery processes that are influenced or dictated by the test owner or test administrator. For example, test

length is often determined by a combination of the test blueprint (i.e., to ensure appropriate content coverage) and the amount of time an examinee will have to complete the test (e.g., classroom time, cost of seat time at a testing center, etc.). In terms of selecting an IRT measurement model, the Rasch model is often preferred due to its simplicity – in terms of sample size required for calibration as well as the fact that scoring is more intuitive and easier to explain to examinees, parents, and other stakeholders.

Use of Multiple Measures

One of the problems with person-fit indices is that a significant statistic only tells the researcher that the model doesn't fit the data; it does not provide the researcher with potential causes for the misfit. Several studies (Karabatsos, 2003; Meijer, 2003; Meijer & Sijtsma, 2001) have explored the sensitivity of various indices to various types of aberrance (e.g., cheating behavior, careless or random responding, etc.). In comparing 36 different person-fit measures, Karabatsos (2003) found that cheating behavior was the most difficult to detect, while careless and random responding were the easiest to detect. If different measures do in fact detect different types of aberrant responding, researchers should consider the combined use of two or more measures to detect and possibly identify the cause of misfit. One example of this is using a scalar measure (e.g., l_z) to flag individual item response vectors showing misfit and then using the Person Response Curve as a diagnostic tool to determine possible causes (Sijtsma & Meijer, 2001).

When it comes to cheating behavior, the implications of classifying a test score as questionable or invalid can be damaging to both the examinee and test administrator. Meijer and Sotaridona (2006) have strongly recommended using multiple measures when flagging examinees as aberrant responders, especially in the case of cheating.

Using a single measure to classify an examinee's test-taking behavior as aberrant is not advisable. Multiple measures can be used to assist in making decisions regarding the fit of an examinee's response pattern to the model. By gathering multiple pieces of evidence of aberrance, the test administrator will be better able to build a case warranting possible action such as requiring the examinee to retake the exam or making some kind of score adjustment.

Item Response Time

Over the past two decades, as computer-based test administration has become more prevalent, research has emerged suggesting the use of item response time as a measure to detect aberrant responding. Computer-based test administration allows for the collection of data regarding how much time an examinee spends on individual items, from the moment the item is presented on the screen to when a response is entered and the examinee moves to the next item. In addition, more precise information regarding the amount of time spent on the test as a whole is available. Many tests allow examinees to move backwards and review or even change responses submitted for previous items. Item response time offers new information about test-taking behavior not previously available with paper-based test administration. In a research study conducted by Hauser, Kingsbury, and Houser (2011), the authors concluded that there was reasonable evidence to "conceptualize inappropriate test-taking behavior as manifest at the item level" (p. 4) with regard to item response time. These findings were corroborated in a study by Wang and Xu (2015), in which the authors utilized a mixture hierarchical model to identify examinees exhibiting rapid guessing test-taking behavior based on item response and response time data.

This section presents methods for incorporating item response time into the identification of aberrant responding.

Response Time Models

As noted by Schnipke and Scrams (1999), response times have had a role, albeit limited, in testing for a long time. Several models of response time have been proposed to explore topics such as speed-accuracy relationships, speeded tests, test-taking strategies (e.g., pacing and guessing behavior), and subgroup differences or fairness (Schnipke & Scrams, 1999). With computer-based testing offering an unobtrusive and convenient way to collect timing data, research has expanded to include response time models focused on the use of response time data in improving item and person parameter estimates and improving the efficiency of item selection rules in computer-adaptive testing (van der Linden, 2009). Finally, there has been considerable research conducted on response time models that focus on examinee motivation or low levels of effort exhibited by an examinee when taking a test (Wise & DeMars, 2006; Wise, Ma, Kingsbury & Hauser, 2010; Wise & Kong, 2005). This type of behavior is exhibited by rapid responding or guessing in responding to test items.

There is limited research available on applying response time models to the detection of aberrant responding such as cheating. Van der Linden and van Krimpen-Stoop (2003) implemented a loglinear model of response time to check for aberrance to complement IRT fit analysis of an item response vector. Specifically, the authors modeled response time as a variable with a lognormal distribution and utilized a series of classical and Bayesian residual checks to assess model fit of individual response time vectors. The study produced mixed results with varied success in detecting simulated aberrance, but also varied levels of false positives.

Effective Response Time

Meijer and Sotaridona (2006) developed a measure of response time to detect examinees with knowledge of the test items prior to actually taking the test (indicating cheating behavior). The measure, called Effective Response Time (ERT), was defined

as “the time required for an individual examinee to answer an item correctly” (p. 1). This measure is of particular interest because it was designed specifically to distinguish examinees who may have item preknowledge (cheaters) from those who do not (non-cheaters) through their response time. The authors demonstrated the effectiveness of the ERT in identifying examinees with simulated preknowledge.

To utilize the Effective Response Time measure, an ERT is computed for each item for each examinee meeting the following requirements:

- the probability of a correct response for an examinee with a given level of ability is greater than chance (i.e., examinee should be able to select the correct answer for the item), and
- the examinee has selected the correct answer for the item.

Unexpectedly short observed item response times (i.e., those that deviate from the ERT) are assumed to be an indication of item preknowledge. The rationale for these requirements lends support for the use of ERT as a means for detecting cheating behavior because they were established to reduce the variability or noise in item response time data caused by other types of response behavior. For example, lower ability examinees may spend less time responding to an item but select the correct answer just by a lucky guess.

To model response time, the authors utilized the methodology for a loglinear model outlined by van der Linden and van Krimpen-Stoop (2003):

$$\ln T_{ij} = \mu + \delta_i + \tau_j + \varepsilon_{ij} , \qquad (7)$$

with

$$\varepsilon_{ij} \sim N(0, \sigma^2) ,$$

where

$\ln T_{ij}$ is the natural logarithm of the response time for examinee *j* for item *i*,

δ_i is the response time required by item *i*,

τ_j is the slowness of examinee *j*,

μ is the general response time for the population of examinees and test items, and

ε_{ij} is a normally distributed residual or interaction term for item *i* and examinee *j* with

mean 0 and variance σ^2.

The parameters for the loglinear response time model can be estimated as follows:

$\mu \equiv E_{ij}(\ln T_{ij})$,

$\delta_i \equiv E_j(\ln T_{ij}) - \mu$,

$\tau_j \equiv E_i(\ln T_{ij}) - \mu$,

$\sigma^2 \equiv E_{ij}(\ln T_{ij} - \delta_i - \tau_j)^2$.

Assuming known item parameters and person ability estimates computed according to a specified IRT model, a set of examinees $j = 1, \ldots, J_i$ is selected such that the probability of a correct response is greater than chance and the correct response was selected for the item. Using the parameters computed for the loglinear response time model, Meijer and Sotaridona proposed estimating the ERT for each item *i* for each examinee *j* by regressing $\ln T_{ij}$ on θ_j and τ_j:

$$\ln T_{ij} = \beta_0 + \beta_1\theta_j + \beta_2\tau_j + \varepsilon_j , \quad (8)$$

where the β's are regression coefficients, ε_j is an error term assumed to be normally distributed with mean 0 and variance σ^2, and the ERT is:

$$\widehat{\ln T_{ij}} = E(\beta_0 + \beta_1\theta_j + \beta_2\tau_j + \varepsilon_j) = \hat{\beta}_0 + \hat{\beta}_1\theta_j + \hat{\beta}_2\tau_j . \quad (9)$$

The observed response time for an examinee suspected of item preknowledge or cheating (*c*) can then be tested against the ERT for that item. Assuming response time is normally distributed on the log scale, we can use the standardized form

$$z_{ic} = \frac{\ln T_{ic} - \widehat{\ln T_{ij}}}{\sigma_i} \quad (10)$$

where $\ln T_{ic}$ is the item response time for an examinee suspected of cheating and σ_i^2 is the variance of the log response time for item *i*:

$$\sigma_i^2 = (J_i - 1)^{-1} \sum_j^{J_i} (\ln T_{ij} - \widehat{\ln T_{ij}})^2 \quad (11)$$

Using the assumption that response time is normally distributed under a log scale, it follows that z_{ic} follows a standard normal distribution and that z_{ic}^2 follows a chi-square distribution with one degree of freedom. Therefore, the sum of z_{ic}^2 also follows a chi-square distribution with the degrees of freedom equaling the number of items in the summation

$$X_c = \sum_i z_{ic}^2 \sim \chi_{ic}^2 .$$

To flag for item preknowledge, the quantity $P(X_c \geq x) = p$ can then be tested against a level of significance α, with values of *p* less than α indicating item preknowledge.

The authors cautioned against using the ERT measure as the only measure used to classify someone as cheating. The authors explained that an examinee who is cheating on an exam may spend a reasonable amount of time on the exam just to avoid suspicion. The ERT was intended to be used as one piece of empirical evidence supporting a claim of aberrant responding.

Research Using Simulated Data

The majority of the research conducted on person fit has utilized simulated data. Simulated data allow the researcher to control the conditions (e.g., type of aberrant behavior exhibited, test-taker characteristics, theta levels, etc.) under investigation, while studies utilizing real data pose significant challenges in that it is difficult to obtain a priori knowledge of examinee motivation and/or behavior. Cheating behavior is especially difficult to detect in practice since researchers typically do not know which test items

may be compromised and therefore, on which items examinees have prior knowledge. In addition, as discussed earlier, detection rates vary across indices and are influenced by factors such as test length and theta level.

Rupp (2013) provided a comprehensive review of simulation studies conducted in person-fit research. He argued that, for the results of simulation studies to be of practical use, design factors should match or be similar to "the kinds of real-life application contexts that practitioners operate in" (p. 12). Rupp went on to delineate questions that should be considered when designing a person-fit simulation study to ensure that the results can be applied to an operational setting. As Rupp gleaned these questions from a thorough review of the person-fit literature, it is not surprising that many are routinely incorporated into the research. The primary questions or considerations can be summarized as follows:

- **Percentage of persons responding aberrantly.** Rupp noted that the percentage of aberrant responders in a sample can impact the ability of indices to identify the aberrant responders. As such, he recommended that this design factor be included in simulation studies. This question could also be interpreted to refer to the percentage of the target population that is expected to respond aberrantly. For example, is estimating the percentage of aberrant responders in a group of examinees who tested over a 12-month period a viable method for determining the ranges or levels of aberrance to simulate?
- **Characteristics of persons responding aberrantly.** This consideration refers to the way aberrance is operationalized within the study. For example, when simulating cheating behavior, it is often assumed that lower ability examinees are more likely to exhibit this type of aberrance. Additionally, it should be apparent how the condition of "low ability" is defined quantitatively.

- **How persons respond aberrantly and to which items.** With regard to the items on an exam, the research must consider the number and type of items examinees are likely to answer aberrantly. For example, are the "hard" or "easy" items more likely to be targeted for cheating behavior? In addition, Rupp noted that the "direction and magnitude of the induced effect" should be explicitly defined (p. 14). For example, he described methods used in research to change non-aberrant responses to aberrant by replacing responses for specific items as either deterministic (e.g., changing incorrect responses to correct for low ability examinees to simulate cheating) or probabilistic (e.g., changing responses from incorrect to correct with a specified probability).

Rupp also provided recommendations for design factors that impact the generalizability of a simulation study. Primary factors he listed include the IRT model used to generate data, sample size, distribution of person and item parameters, number of items or length of the test, number of replications, and Type I error rate. He stressed that, while most of these factors are considered and addressed in simulation studies, researchers must take care to provide thorough and detailed documentation of how each of these factors was operationalized and implemented within a study to support generalization and allow for practical use of outcomes.

A predominant topic in the literature regarding the design and use of simulation studies to evaluate the effectiveness of person-fit indices is regarding hypothesis testing (Nering, 1995; Seo & Weiss, 2013; van Krimpen-Stoop & Meijer, 1999). As discussed above, there are many factors that impact the distribution of a person-fit statistic and therefore, the use of hypothesis testing to classify a response pattern as aberrant.

In addition, many researchers have stressed the importance of the practical application of research results (Meijer, 2003; Nering, 1995; Rupp, 2013). As such, an important focus of several studies has been on confirmation of the theoretical sampling

distribution of a person-fit index to allow for hypothesis testing. The most promising results are from studies in which real data are used to estimate item and person parameters (Seo & Weiss, 2013; van Krimpen-Stoop & Meijer, 1999). In fact, Seo and Weiss (2013) recommended that "Monte Carlo simulation be implemented using the item parameters estimated with real data and an appropriately modeled distribution of θ" (p. 1014) to determine the observed person-fit distribution and appropriate critical values for identifying aberrance.

Research Questions

The primary purpose of this research was to determine whether using multiple IRT-based person-fit indices in the identification of a specific form of aberrant responding, namely cheating behavior, on an exam is more effective than relying on a single measure. Cheating behavior in this context is defined as item pre-knowledge, or simply stated, when an examinee has obtained knowledge of item content prior to taking the exam.

The goal of a multiple-measure approach is to gather evidence from various points of view to either support or dispel an assertion of cheating. This research explored whether using indices that evaluate different aspects of a response pattern (likelihood of the response pattern, response time, shape of the curve) provides complementary information in the assessment of aberrant responding, thus increasing the accuracy of detection rates.

A secondary purpose of this research was to evaluate the impact of various factors on the effectiveness of person-fit indices in identifying aberrant response patterns. The factors (independent variables) that were evaluated included IRT model, exam length, ability level, degree of aberrance within a response vector, and degree of aberrance within a given sample of examinees. The dependent variables were person-

fit Type I error rate at α = .05 and the classification accuracy of aberrant responding using single or multiple measures.

The research questions addressed by this study included:

1. How does the degree of aberrance exhibited within a response pattern and an examinee population impact the effectiveness of indices in identifying cheating behavior for short and long tests and for examinees with low or mid-range ability levels?
2. How does IRT model selection (Rasch versus two-parameter logistic) impact the effectiveness of indices in identifying cheating behavior?
3. How does the use of multiple person-fit indices (scalar, timing, graphical) affect the identification of aberrant responding, namely cheating behavior?

CHAPTER 3: METHODS

A simulation study was conducted to evaluate the impact of various factors on the effectiveness of three person-fit indices, alone and in combination, in the identification of cheating behavior. The factors examined included:

- IRT model: Rasch; 2PL
- Exam length: *Short form* (40 items); *Long form* (100 items)
- Theta level: The theta level of the examinees exhibiting cheating behavior – *low level* (theta values in the bottom 30%); *mid level* (theta values in the middle 30%).
- Percentage of exam aberrance: The specified percentage of item responses and response times manipulated to represent cheating behavior – 10%, 25%, and 50%.
- Percentage of sample aberrance: The specified percentage of examinees exhibiting cheating behavior within a given theta level – 5%, 15%.

The Rasch model is one of the most commonly used IRT models in operational exam programs as it is mathematically less complex and therefore easier to implement and has fewer problems with estimating parameters. In addition, raw score (i.e., number correct) is a sufficient statistic so every examinee with the same raw score on an exam will receive the same ability estimate. This allows for easier interpretation by test users and stakeholders (e.g., examinees, parents, educators). One assumption of the Rasch model is that all items on an exam have the same level of discrimination ability (i.e., a = 1.00). In practice, however, this assumption may not be reasonable. Recently, the two-parameter logistic (2PL) model has gained in popularity because it does allow item discrimination to vary, but does not have the same problems with estimation of the

guessing parameter as does the three-parameter logistic model. In order to conduct a real-parameter simulation study (as described below), item and person parameters from an exam program that currently utilizes the Rasch model were employed. The 2PL model was selected to evaluate whether IRT model has an impact on person-fit indices because it introduces an additional parameter (i.e., *a*-parameter) to the measurement model and is a logical choice for practitioners exploring alternatives to the Rasch model.

The short and long exam lengths represent the lower and higher ends of the range of total number of items generally seen on operational or commercial exams. For the remaining factors (theta level, percentage of exam aberrance, percentage of sample aberrance), determining values that would be representative of examinees likely to cheat on exams proved more difficulty as these groups do not generally admit to cheating.

With regard to the ability levels used for this research, the low-level ability examinees were defined as those with theta estimates in the lower 30% of the sample and mid-level ability examinees were those with theta estimates in the middle 30% of the sample. These definitions are similar to the quintiles used by Drasgow, Levine, and Williams (1985) to classify ability levels of SAT examinees as low, moderately low, and average in their research on aberrant responding. Quintiles were also used to categorize GRE examinees by ability level, with the first three quintiles labeled as low, low middle, and middle (Schaeffer, Reese, Steffen, McKinley, & Mills, 1993).

A review of the literature provided levels of exam aberrance at which various person-fit indices are effective. The level of exam aberrance represents the percentage of exam items that an examinee has been previously exposed to and recalls when taking the exam. For example, an examinee may have obtained a list of item content (including the question and correct response option) harvested by previous examinees. While the list may not contain all the items on the new examinee's test, it does contain some

percentage of the items. The exam aberrance levels (10%, 25%, 50%) used for this study were based on research conducted by St-Onge et al (2011).

The prevalence of cheating within a given population of examinees proved the most difficulty to obtain. According to Cizek (1999), determining the frequency of cheating on exams is an “apples-and oranges endeavor” (p. 14). This is due in part to the various definitions of cheating used by researchers and the various methods used to collect data. In addition, professional certification or licensing organizations and higher education institutions often do not want this type of information disclosed as it could jeopardize the reputation and integrity of the exam program and create legal ramifications. In one study attempting to summarize the prevalence of cheating among college students, it was reported that 27% - 37% of students used advance knowledge of test content to cheat on an exam during their college career (Cizek, 1999, p. 27). Because higher-stakes testing programs generally have higher levels of security in place to deter cheating, it was assumed that the frequency of cheating on a given exam during a given administration would be on the lower end of the spectrum (5% and 15%) and limited to examinees with low to middle level ability, assuming they would have more motivation to cheat in order to improve their test score.

In summary, data were generated to simulate cheating behavior under each factor (IRT model *by* exam length *by* theta level *by* degree of exam aberrance *by* degree of sample aberrance). All factors were crossed, producing a total of 48 (2 x 2 x 3 x 2 x 2) simulated conditions. The first dependent variable examined was the effectiveness of person-fit measures (alone and in combination) in correctly identifying aberrant responding (Type I error rate at $\alpha = .05$). The second dependent variable was the classification accuracy of aberrant responding using individual or combined (multiple) measures.

Real-Parameter Simulation

A simulation approach, referred to as real-parameter simulation, was proposed by van Krimpen-Stoop and Meijer (1999) for use in person-fit research. Under the real-parameter simulation approach, response data are simulated using the item parameters estimated from real examinee response data and modeled using the estimated theta distribution from the real examinees. Additionally, Seo and Weiss (2013) demonstrated that this approach is effective for generating datasets to determine the critical values for the person-fit indices being researched.

A variation of the real-parameter simulation approach (Seo & Weiss, 2013) was utilized to simulate item response and response time data for this research. Specifically, Monte Carlo simulation was implemented using exam information (item parameters, theta values, response time parameters) from a language skills exam developed for use as part of the admissions process for an international graduate school program.

The language skills operational exam forms were constructed by selecting a set of items from a large pool of items to meet both content and statistical specifications. Each form consisted of 32 dichotomously-scored, multiple-choice items. The items contained five options and required selection of the single, best response. A concurrent calibration was performed during the development phase of the exam program to estimate item parameters and place all items on a common scale. Items were calibrated using WINSTEPS® Rasch measurement software (Linacre, 2016), with ability estimates computed using the joint maximum likelihood estimation (JMLE) procedure. New items are continually pretested on operational exams in order to increase the size of the operational item pool and refresh the exam forms.

Response Time Data

The language skills exam is a computer-based test, which allows response time data to be collected during administration. Specifically, the amount of time required for

an examinee to respond to each item is captured, which allows the computation of both item and person response time parameters. In addition to simulating item response data, response time data were simulated using the loglinear model outlined by van der Linden and van Krimpen-Stoop (2003) and provided in Equation 7. Researchers have found that this model provides simulated response time data that are a good fit to actual response time data (Schnipke & Scrams, 1999; Wang, Xu, & Shang, 2016).

The language skills response time data were used to estimate the parameters for the loglinear response time model (van der Linden & van Krimpen-Stoop, 2003). As shown in Equation 7, these parameters included the (a) response time (δ) for each item, (b) slowness factor (τ) for each examinee, (c) overall response time for examinees and items (μ), and (d) error term (ε) or residual for each examinee-by-item interaction.

IRT Model

The real item and person parameters from the language skills exam were used to simulate data under the Rasch and 2PL models. In Hambleton and Swaminathan (1985), the mathematical form of the logistic model is given by

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}}, \qquad (12)$$

where

$P_i(\theta)$ = the probability that examinee with ability level θ answers item i correctly,

b_i = the item difficulty parameter,

a_i = the item discrimination parameter, and

D = 1.7 (scaling factor).

Note that in using the general form of the IRT logistical model in Equation 12, with the scaling factor (D) set to 1.7, a_i is set to 0.58823 for all items in order to fix the item discrimination parameter at a value of 1.00 to implement the Rasch model.

To simulate data under the 2PL model, the real item response data from the language skills exam forms were re-calibrated under the 2PL model using flexMIRT® (Cai, 2013). A linear transformation using the Mean/Sigma method (Kolen & Brennan, 2014) was then applied to the 2PL location parameters to put them on the same scale as the Rasch parameters in order to minimize differences in item, and therefore form, difficulty due to calibration method. Note that item location is referred to as item difficulty or *b*-parameter and item discrimination is referred to as item discrimination or *a*-parameter going forward.

Since one of the goals of this research was to evaluate whether selection of IRT model (Rasch versus 2PL) had an impact on the effectiveness of indices in identifying cheating behavior, the theta (θ) values corresponding to response vectors were not re-estimated using the 2PL item parameters. That is, the student theta values used for the Rasch simulations were also used for 2PL simulations to avoid any confounding effects that may have been introduced by re-estimation.

Data Generation

Data sets were generated in a two-step process. During the first step, Monte Carlo simulation procedures were utilized to generate item response and response time data for each IRT model and exam length combination using SAS® statistical analysis software. The second step of the process was to manipulate the data in order to create each of the 48 study conditions.

Step 1: Baseline Data

To create the short and long exam forms for this study, items in the calibrated language skills item pool were first ordered by difficulty (*b*-parameter) and categorized by quintile. The appropriate number of items per form (40 for the short form and 100 for the long form) were then randomly drawn from the pool by quintile so that the *b*-parameters for a form approximated a normal distribution, $N(0.7, 1.2)$. Items were

selected from each quintile proportionately so that the overall form difficulties would be comparable. In addition, the same randomization seed was used to select items for the 40-item and 100-item forms. The associated item parameters (*b*-parameters for the Rasch and 2PL conditions; *a*-parameters for the 2PL condition) and item response time parameters (δ) were used in the data simulation. Tables 2 and 3 provide summaries of item statistics for the 40-item and 100-item forms, respectively, based on the real data.

Table 2. *Summary Item Statistics for the 40-item Forms*

Model	Variable	Mean	SD	Min	Max
Rasch	*b*-parameter	0.573	1.011	-1.941	2.106
2PL	*a*-parameter	0.778	0.324	0.328	1.903
	b-parameter	0.558	0.970	-2.371	1.952
	Item time[1]	40.218	16.781	15.675	73.537
	Delta	-0.010	0.393	-0.778	0.717

Note: *a*-parameter under Rasch model is effectively fixed at 1.000 for all items
[1]Item time is reported in seconds

Table 3. *Summary Item Statistics for the 100-item Forms*

Model	Variable	Mean	SD	Min	Max
Rasch	*b*-parameter	0.520	1.043	-2.117	2.790
2PL	*a*-parameter	0.766	0.319	0.309	1.903
	b-parameter	0.514	1.014	-2.371	2.497
	Item time[1]	41.920	16.457	15.675	73.537
	Delta	0.024	0.385	-0.779	0.717

Note: *a*-parameter under Rasch model is effectively fixed at 1.000 for all items
[1]Item time is reported in seconds

Theta values used in the simulation were drawn from a pool of over 70,000 examinees that completed the language skills exam. Theta values (N = 1,000) were randomly selected according to a normal distribution, $N(-0.4, 1.1)$. The theta (θ) values and corresponding person time or tau (τ) parameters were used to simulate item response data for both the short and long forms across replications. Table 4 provides a summary of the person statistics (N = 1,000) based on the real data.

Table 4: *Summary of Real Data Person Statistics*

Variable	Mean	SD	Min	Max
Theta	-0.332	1.065	-3.687	3.412
Tao	-0.066	0.211	-1.338	0.248
Average item response time[1]	39.173	3.015	15.406	40.813

Note. N = 1,000.

[1]Average item response time is reported in seconds

Step 2: Cheating Behavior Simulation

A baseline sample of data (*N* = 1,000) was simulated for each IRT model and exam length assuming no cheating behavior present (i.e., no item response or response time data were manipulated). The baseline data were then manipulated to create samples of examinees exhibiting aberrant behavior following each research condition. Table 5 shows the simulated conditions for each cell of the research design. The procedures used to mimic cheating behavior were similar to those followed by Meijer and Sotaridona (2006) and are described in detail below. As noted above, Rupp (2013) classified this type of data manipulation as deterministic (e.g., changing incorrect responses to correct for a specified number of low ability examinees). In general, items with difficulty values greater than the examinee's ability level and incorrect responses were randomly selected. Incorrect item responses were then changed to correct and response times were reduced by 50%. This process was continued for varying percentages of items to meet each study condition.

Table 5: *Simulated Cheating Conditions*

			Theta level: Low			Theta level: Mid		
			Exam aberrance			Exam aberrance		
Model	EL	SA	10%	25%	50%	10%	25%	50%
Rasch	Short	5%						
		15%						
	Long	5%						
		15%						
2PL	Short	5%						
		15%						
	Long	5%						
		15%						

Note: Baseline *N* = 1,000 for each IRT model by exam length. EL = exam length. SA = sample aberrance.

For each cell within a given IRT model and form length, data were manipulated as follows:

1. Randomly select a case within the specified *theta level*
    a. Randomly select an item (without replacement).
    b. If *b*-value > θ and item response is incorrect (value = 0)
        i. change item response to correct (value = 1);
        ii. reduce item response time by 50%.
    c. Repeat steps a and b until the specified percentage of *exam* aberrance is achieved.
2. Repeat steps 1a-c until the specified percentage of *sample* aberrance is achieved.

After manipulating the data to mimic cheating behavior, θ and τ values for examinees were re-estimated using the altered response vector. It should be noted that the item parameters (*a, b,* δ) were not re-estimated after manipulating the data for aberrance. This is representative of a typical operational exam program, in which the

item parameters are fixed after the pilot or pretesting data are collected and final calibration is performed.

When three replications of baseline data for each IRT model had been generated and manipulated to simulate cheating, the data were checked to ensure that all procedures were being carried out as intended. Verifications included ensuring that the appropriate percentages of exam and sample aberrance were present across exam length and theta level.

Additional steps were taken to verify the entire simulation. Correlations between the original ("true") theta estimate and raw score (total number correct) over conditions were examined. As expected, the correlation coefficients go down as the amount of cheating behavior within response vector increases. Correlations between the person time (τ) and the average person response time over conditions were also examined and the same pattern was exhibited, namely a decrease in correlation coefficients as cheating increased. Results of correlation analyses are presented in Tables 6 and 7. Note that correlations between theta and raw score for the baseline data (i.e., no aberrance) were 0.928 and 0.964 for the Rasch model, short and long forms, respectively, and 0.941 and 0.970 for the 2PL model, short and long forms, respectively.

Table 6: *Correlation Between Original Theta and Raw Score by Condition*

			Theta level: Low			Theta level: Mid		
			Exam aberrance			Exam aberrance		
Model	EL	SA	10%	25%	50%	10%	25%	50%
Rasch	Short	5%	0.915	0.872	0.747	0.924	0.897	0.813
		15%	0.894	0.768	0.441	0.919	0.853	0.691
	Long	5%	0.953	0.909	0.776	0.959	0.929	0.836
		15%	0.934	0.808	0.461	0.953	0.880	0.702
2PL	Short	5%	0.932	0.896	0.788	0.939	0.916	0.841
		15%	0.915	0.812	0.517	0.935	0.878	0.729
	Long	5%	0.961	0.926	0.814	0.966	0.941	0.860
		15%	0.948	0.847	0.540	0.961	0.898	0.738

Note: EL = exam length. SA = sample aberrance.

Correlations between tau and person response time for the baseline data (i.e., no aberrance) were 0.866 and 0.936 for the Rasch model, short and long forms, respectively, and 0.866 and 0.936 for the 2PL model, short and long forms, respectively.

Table 7: *Correlation Between Original Tau (τ) and Average Person Response Time by Condition*

			Theta level: Low			Theta level: Mid		
			Exam aberrance			Exam aberrance		
Model	EL	SA	10%	25%	50%	10%	25%	50%
Rasch	Short	5%	0.860	0.824	0.724	0.856	0.812	0.693
		15%	0.852	0.775	0.622	0.841	0.737	0.539
	Long	5%	0.927	0.879	0.757	0.924	0.868	0.727
		15%	0.913	0.814	0.637	0.904	0.778	0.555
2PL	Short	5%	0.860	0.824	0.724	0.857	0.812	0.693
		15%	0.852	0.775	0.622	0.841	0.737	0.540
	Long	5%	0.927	0.879	0.757	0.924	0.868	0.728
		15%	0.913	0.814	0.637	0.905	0.778	0.553

Note: EL = exam length. SA = sample aberrance.

To evaluate the data at the item level, summaries of the difference between the original item *p*-value and the *p*-value computed within the cheating conditions were examined. The results show that items appear to be getting easier (i.e., *p*-values are increasing) as cheating behavior increases. Similarly, the difference between the average item time from the baseline data and the average item time within the cheating conditions decreases as cheating behavior increases. Summary data can be found in Tables 8 and 9.

Table 8: *Average Difference Between p-Values Before and After Data Manipulation*

			Theta level: Low			Theta level: Mid		
			Exam aberrance			Exam aberrance		
Model	EL	SA	10%	25%	50%	10%	25%	50%
Rasch	Short	5%	0.005	0.0125	0.025	0.005	0.0125	0.025
		15%	0.015	0.0375	0.075	0.015	0.0375	0.075
	Long	5%	0.005	0.0125	0.025	0.005	0.0125	0.025
		15%	0.015	0.0375	0.075	0.015	0.0375	0.075
2PL	Short	5%	0.005	0.0125	0.025	0.005	0.0125	0.025
		15%	0.015	0.0375	0.075	0.015	0.0375	0.075
	Long	5%	0.005	0.0125	0.025	0.005	0.0125	0.025
		15%	0.015	0.0375	0.075	0.015	0.0375	0.075

Note: EL = exam length. SA = sample aberrance.

Table 9: *Average Difference Between Item Time Before and After Data Manipulation*

			Theta level: Low			Theta level: Mid		
			Exam aberrance			Exam aberrance		
Model	EL	SA	10%	25%	50%	10%	25%	50%
Rasch	Short	5%	-0.008	-0.020	-0.040	-0.008	-0.020	-0.041
		15%	-0.024	-0.059	-0.119	-0.024	-0.061	-0.122
	Long	5%	-0.008	-0.020	-0.040	-0.008	-0.021	-0.041
		15%	-0.024	-0.060	-0.121	-0.025	-0.062	-0.123
2PL	Short	5%	-0.008	-0.020	-0.040	-0.008	-0.020	-0.041
		15%	-0.024	-0.059	-0.119	-0.024	-0.061	-0.122
	Long	5%	-0.008	-0.020	-0.040	-0.008	-0.021	-0.041
		15%	-0.024	-0.060	-0.121	-0.025	-0.062	-0.123

Note: EL = exam length. SA = sample aberrance.

To allow for visual inspection of a baseline PRC graph compared to the PRC graph for the corresponding manipulated data record, one set of PRC graphs for each study condition was generated. Figures 1 and 2 show a set of PRCs for study conditions Rasch x short form x low-level ability x 5% sample aberrance x 10% exam aberrance before data were manipulated (original, baseline data) and after data were manipulated, respectively.

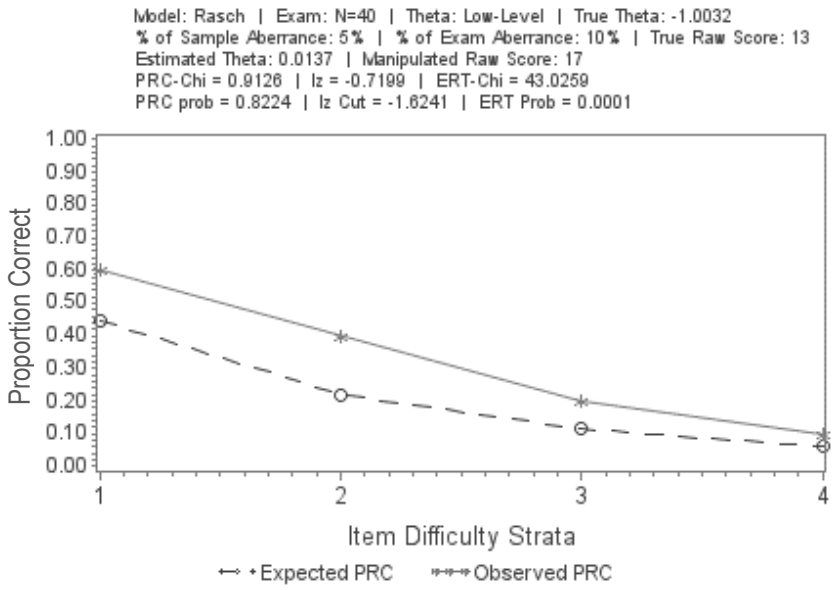

Figure 1. Baseline PRC for condition: Rasch model x short form x low-level ability.

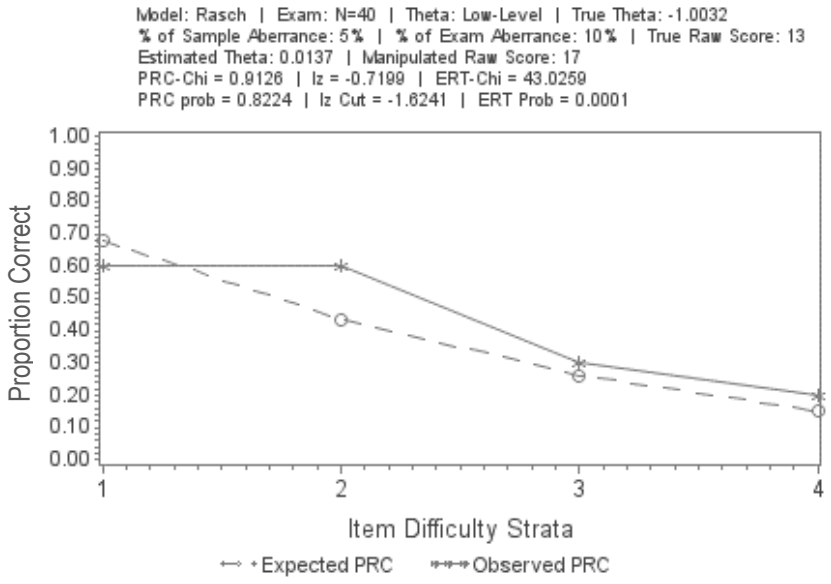

Figure 2. PRC for cheating condition: Rasch model x short form x low-level ability x 5% sample aberrance x 10% exam aberrance.

Figures 3 and 4 show a set of PRCs for study conditions 2PL x long exam x mid-level ability x 15% sample aberrance x 50% exam aberrance before data were manipulated (original, baseline data) and after data were manipulated, respectively.

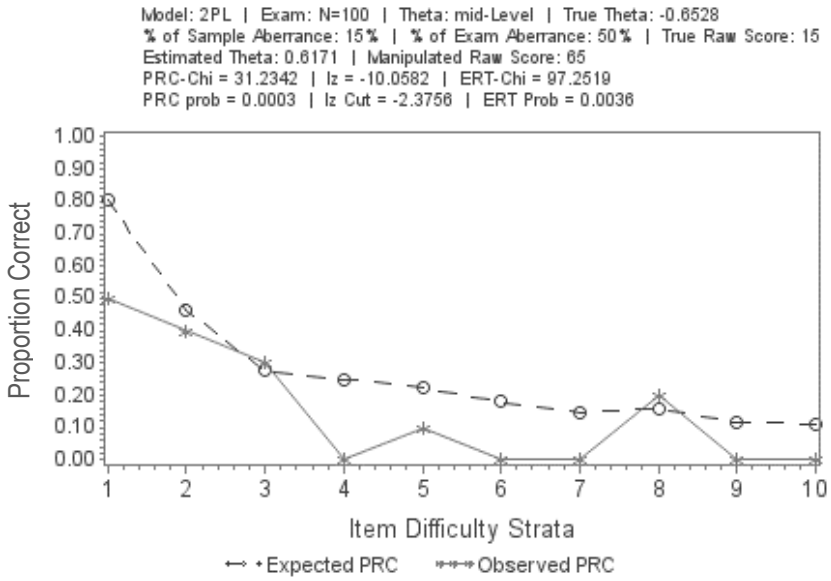

Figure 3. Baseline PRC for condition: 2PL model x long form x mid-level ability.

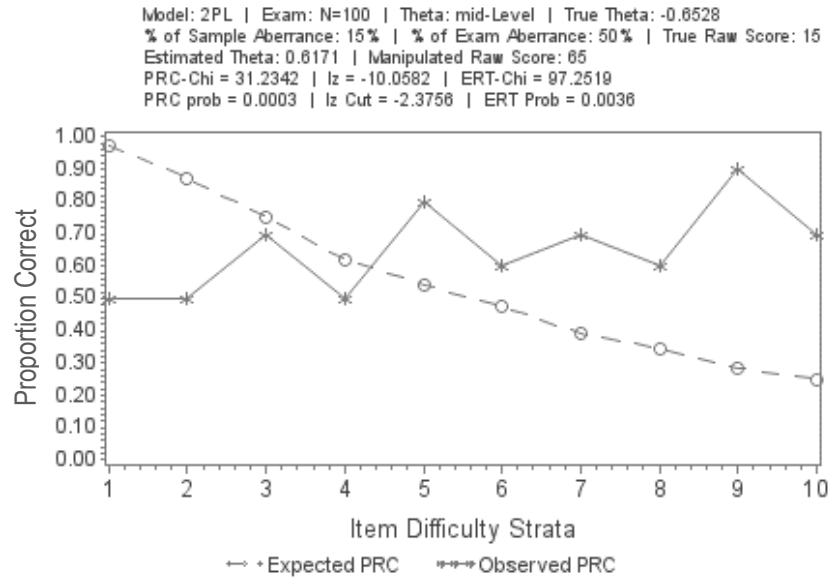

Figure 4. PRC for cheating condition: 2PL model x long form x mid-level ability x 15% sample aberrance x 50% exam aberrance.

By comparing Figures 1 and 3 with Figures 2 and 4, respectively, it can be seen that item responses were changed from incorrect to correct. Appendix A contains PRCs representing the remaining study conditions.

Replications. As noted in Harwell, Stone, Hsu and Kirisci (1996), the number of replications in a simulation study can be viewed as analogous to sample size in an empirical study and therefore, should be determined using the same criteria (p. 110). To determine the number of replications required to provide adequate estimates of power of the ANOVA *F* test and to detect meaningful differences between conditions, the procedures described by Harwell et al. (1996) were utilized. Specifically, replications were generated in increments of 100. After each 100 replications, a factorial ANOVA was conducted for each person-fit measure using Type I error rate as the dependent variable. Effect size *f* was estimated for each study condition using eta squared (η^2), which is computed using the sums of squares for a given factor (SS_{effect}) and the sums of squares for the total model (SS_{total}). That is,

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}} \tag{13}$$

and

$$f = \sqrt{\frac{\eta^2}{(1-\eta^2)}}\,. \tag{14}$$

The estimated effect size *f*, number of replications (*N*), *df* for the effect, and the total number of cells in the model were then used to estimate power at $\alpha = .05$ in the software program G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007).

Power estimates were computed after 100, 200, 300, 400, and 500 replications. Power estimates were mostly stable and high (power >= .997) across person-fit measures and study factors as the number of replications increased. The sample aberrance study factor, however, along with several interaction effects that included

sample aberrance, showed low power estimates across replications. Because this research included a total of 48 study conditions, it was determined that a higher number of replications would be utilized. Therefore, results discussed in the remaining sections and chapters are based on data obtained from 500 replications.

Person-Fit Measures

The measures used to identify cheating behavior included the following:

- l_z (likelihood-based, scalar measure)
- Person Response Curve (graphical measure) and χ^2 (residual or goodness-of-fit measure)
- Effective Response Time (timing measure)

l_z and Person Response Curve χ^2

The software program WPerfit (Ferrando & Lorenzo, 2000) can be used to compute the person-fit measure l_z (Drasgow, Levine & Williams, 1985) and Person Response Curve χ^2 (Trabin & Weiss, 1983). In addition, WPerfit can produce a graph of the Person Response Curve (PRC) for each examinee. However, there are limitations to using this software with large and numerous data sets (e.g., software cannot be run in batch mode). Therefore, SAS® was used to compute the l_z and Person Response Curve χ^2 statistics using the same equations as those used by WPerfit. This allowed WPerfit to be used to validate the output from SAS®.

To compute l_z, l_o was computed as proposed by Levine and Rubin (1979) and standardized by Drasgow et al. (1985). These equations are consistent with those presented in Equations 1 and 2. Critical values for evaluating a Type I error at $\alpha = .05$ were determined from the l_z distributions computed using the baseline simulation datasets (i.e., datasets with no specified percentage of aberrance). When the data fit the model, values of l_z should be near zero and negative values are an indication of an

inconsistent response pattern (Reise, 1990; de Ayala, 2009). The average critical values by IRT model and exam length are presented in Table 10 below.

Table 10: *Empirical Critical l_z values by IRT Model and Exam Length*

			Critical Value			
Model	EL	*N*[a]	Min	Max	Mean	SD
Rasch	Short	500	-1.821	-1.451	-1.624	0.063
	Long	500	-2.686	-2.381	-2.547	0.048
2PL	Short	500	-1.903	-1.500	-1.684	0.068
	Long	500	-2.560	-2.173	-2.376	0.061

Note: EL = exam length.
[a]*N* represents the number of simulated baseline datasets (i.e., replications)

To compute and graph the observed and expected PRC and also compute the chi-square goodness-of-fit-statistic (PRC χ^2), the procedures provided by Trabin and Weiss (1983) were followed. Specifically, items were ordered by difficulty and divided into groups (called strata). The expected PRC was then constructed from the estimated probability of a correct response for each item (according to the specified IRT model) and averaging the probabilities within strata. The observed PRC was obtained by plotting the proportion of items answered correctly as a function of item difficulty and estimated ability level. Finally, the PRC χ^2 statistic, obtained using the expected and observed average probabilities of correct responses for each strata, was evaluated at $\alpha = .05$ with the degrees of freedom equal to the number of strata minus 1.

Effective Response Time

The Effective Response Time (ERT) measure utilizes items within a response string that meet the following requirements:

- the probability of a correct response is greater than chance based on the estimated theta, and
- the item has a correct response.

Using the parameters computed under the loglinear response time model and following Equations 8 – 11, the ERT for an examinee was computed as the sum of the squared standardized expected response times for each item identified above. SAS® was also used to compute ERT.

Meijer and Sotaridona (2006) stated that, under the assumption that response time is normally distributed under a log scale, it follows that ERT forms a chi-square distribution with the degrees of freedom equal to the number of items in the summation. Therefore, the ERT χ^2 was evaluated at $\alpha = .05$ for the appropriate degrees of freedom.

Analysis

After computing the person-fit measures for simulated examinees within each condition, analyses were performed to address the research questions stated above. The effectiveness of the person-fit measures in detecting simulated cheating behavior, individually and in combination, was examined.

Analysis of Variance (ANOVA)

The Type I error rate of person-fit measures rejecting the null hypothesis (H_0: the model fits the data; response pattern is not aberrant) was the primary dependent variable in this study. The sums of squares (SS) and effect sizes for study factors on Type I error rates were obtained by estimating a factorial ANOVA model in which all between-subject factors were specified. The *F* statistic was evaluated at $\alpha < .05$ for the appropriate degrees of freedom (*df*) to test for statistically significant main effects and interactions.

Indices of effect size computed were η^2 (Equation 13) and *f* (Equation 14). η^2 represents the amount of the total variance that is accounted for by a given effect and provides an indication of the strength of the association between the dependent variable and one or more independent variables. The value of η^2 is interpreted as the proportion of variance accounted for by the effect (Ferguson, 2009; Fritz, Morris & Richler, 2012).

While understanding the meaning of an η^2 is relatively straightforward, it is also beneficial to evaluate the magnitude of the effect against industry convention. Cohen (1992) provides the following guidelines for the interpretation of an effect size given by *f*:

- small $f = 0.10$
- medium $f = 0.25$
- large $f = 0.40$

Classification Accuracy

The value of using multiple measures to identify cheating behavior was of primary interest in this research. To further explore whether there is incremental value and diagnostic efficiency in using more than one person-fit measure, analyses were conducted to evaluate classification accuracy, sensitivity, and specificity.

Cohen's kappa. The degree to which individual person-fit measures agreed in terms of flagging an examinee as cheating was computed using Cohen's kappa (κ) (Cohen, 1960). Coefficient κ can be interpreted as the proportion of classification agreement (or classification accuracy), corrected for chance, where values of 1.00 indicate perfect agreement. Cohen (1960) provides the following equation for computing coefficient κ.

$$\kappa = \frac{p_o - p_c}{1 - p_c}, \tag{15}$$

where

p_o = proportion of agreement,

p_c = proportion of agreement expected by chance, which is computed as the joint probability of the margins.

An approximation of the standard error of coefficient κ is given by

$$\sigma_\kappa = \sqrt{\frac{p_o(1-p_o)}{N(1-p_c)^2}}. \tag{16}$$

As noted by Cohen (1960) and Davenport and El-Sanhurry (1991), the value of κ is constrained by the marginal proportions and values of unity (-1.00 and +1.00) are typically not attainable. Dividing κ by the maximum value of κ allowed by the marginals was provided by Cohen (1960) as a correction to make the values of -1.00 and +1.00 attainable. While this correction was recommended by several researchers (Davenport & El-Sanhurry, 1991), Cohen did not suggest that the corrected κ be used in place of κ. Therefore, the uncorrected κ coefficient was utilized in this study.

The proportion of classification agreement (i.e., uncorrected κ) was examined between individual person-fit measures as follows: l_z x ERT χ^2, l_z x PRC χ^2, ERT χ^2 x PRC χ^2.

Sensitivity and Specificity. Diagnostic efficiency refers to the ability of a measure (or combination of measures) to accurately discriminate between conditions (Doyle, Biederman, Seidman, Weber, & Faraone, 2000; Šimundić, 2008). For this study, diagnostic efficiency is defined as the ability of one or more person-fit measures to accurately discriminate between examinees exhibiting cheating behavior and those not exhibiting cheating behavior. Sensitivity and specificity are two indicators of diagnostic efficiency and can be conceptualized through the use of two-by-two (2x2) contingency tables. For illustrative purposes, an example of a 2x2 contingency table is provided in Table 11.

Table 11: *Diagnostic Efficiency Contingency Table*

		Simulated aberrance	
		Yes	No
Positive person-fit measure(s)	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

Sensitivity refers to the ability of a measure (or combination of measures) to correctly classify examinees who cheated on an exam. Within this study, sensitivity was defined as the proportion of examinees manipulated to exhibit cheating behavior and correctly classified as aberrant (TP/(TP+FN)). Specificity refers to the ability of a measure (or combination of measures) to correctly classify examinees who did not cheat on an exam. Specificity was defined as the proportion of cases that were *not* manipulated to exhibit cheating behavior and *not* classified as aberrant (TN/(TN+FP)).

CHAPTER 4: RESULTS

Type I Error Rate for Individual Measures

Tables containing the full set of results for the factorial ANOVA evaluating the impact of study factors on Type I error rates for the individual person-fit measures are provided in Appendix B. Results presented in this section include only those where the estimated effect size was at least small ($f \geq 0.10$).

The sum of squares (SS) and estimated effect size (f) for interaction effects showing at least a small effect size (bolded) for one or more person-fit measures are provided in Table 12. Overall, the interaction effects produced small or no effect size across the individual person-fit measures. Not surprising, effect sizes (though small) were seen for l_z under the following conditions: EL x EA (f =0.10), EA x T (f = 0.13), EL x EA x T (f = 0.14). The only other notable effect size produced for an interaction effect was for ERT under the condition EA x T, with f = 0.12.

Table 12: *Sums of Squares (SS) and Effect Sizes (f) for Individual Measures Yielding a Small to Large f for Interaction Effects*

	l_z		ERT		PRC	
Effect	SS	*f*	SS	*f*	SS	*f*
EL x EA	11.835	**0.10**	18.484	0.06	57.799	0.07
EA x T	22.348	**0.13**	78.446	**0.12**	54.489	0.07
EL x EA x T	22.744	**0.14**	10.727	0.04	5.944	0.02

Tables 13 - 15 provide the average Type I error rates for the measures yielding small effect sizes under a given interaction.

Table 13: *l_z Average Type I Error Rate for EL x EA Interaction*

EL	EA	l_z
Short	10%	0.405
	25%	0.129
	50%	0.212
Long	10%	0.195
	25%	0.017
	50%	0.009

Note: Standard error was 0.001 for all average Type I Error rates.
EL = exam length. EA = exam aberrance.

Table 14: *l_z and ERT Average Type I Error Rate for EA x T Interaction*

EA	T	l_z	ERT
10%	Low	0.198	0.811
	Mid	0.402	0.580
25%	Low	0.042	0.183
	Mid	0.103	0.138
50%	Low	0.063	0.376
	Mid	0.158	0.419

Note: Standard error was 0.001 for all average Type I Error rates.
EA = exam aberrance. T = Theta level.

Table 15: *l_z Average Type I Error Rate for EL x EA x T Interaction*

EL	EA	T	l_z
Short	10%	Low	0.332
		Mid	0.478
	25%	Low	0.075
		Mid	0.183
	50%	Low	0.121
		Mid	0.304
Long	10%	Low	0.064
		Mid	0.326
	25%	Low	0.009
		Mid	0.024
	50%	Low	0.006
		Mid	0.013

Note: Standard error was 0.001 for all average Type I Error rates.
EL = exam length. EA = exam aberrance. T = theta level.

Simple effects analysis was performed to further explore the two-way and three-way interaction effects for l_z and ERT Type I error rates. The results of the analysis for EL x EA with l_z showed statistically significant differences in EA means for both the short ($F(2, 23952) = 55704.217$, $p < .001$) and long ($F(2, 23952) = 30803.247$, $p < .001$) EL conditions. Simple pairwise comparisons further revealed statistically significant mean differences at $p < .001$ between all levels of EA within both the short and long EL conditions. Figure 5 below provides a graphical representation of the average Type I error rate for l_z across EA and EL. As shown below, there is more variation between the Type I error rates across EA levels for the short EL. Additionally, the difference between Type I error rates under the 25% and 50% EA, long EL conditions appears to be quite small (despite the statistical significant finding), however these were also the smallest Type I error rates within both the short and long EL conditions.

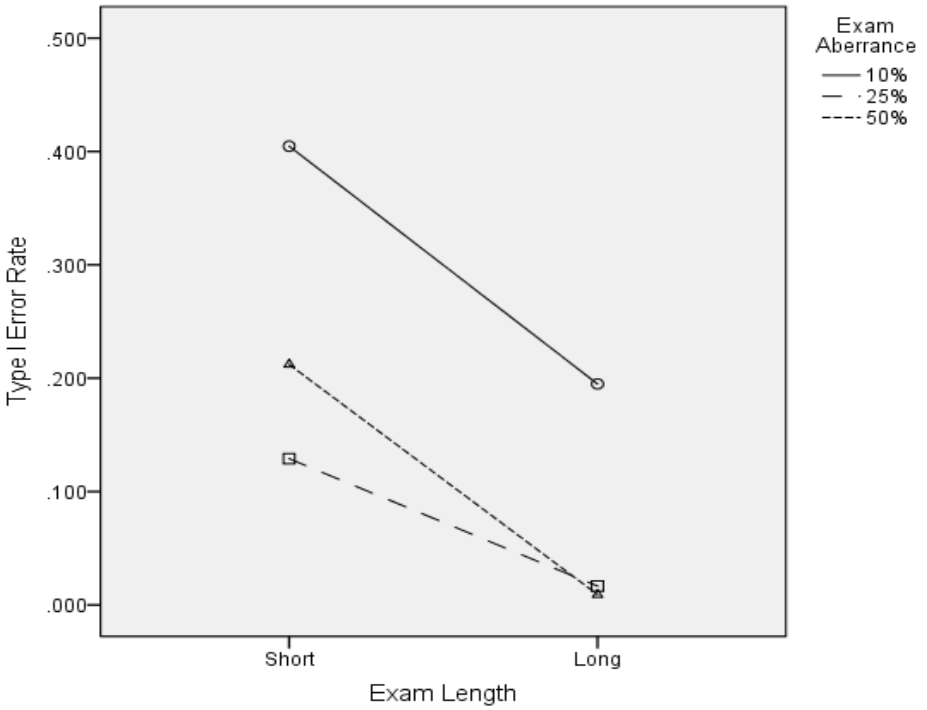

Figure 5. Average Type I error rates for l_z under condition EL x EA.

A simple effects analysis for EA x T with l_z showed statistically significant differences between EA means for both the low ($F(2, 23952) = 19829.914$, $p < .001$) and mid ($F(2, 23952) = 70337.498$, $p < .001$) T levels. Simple pairwise comparisons revealed statistically significant mean differences at $p < .001$ between all levels of EA within both the low and mid T levels. Figure 6 provides graphical representation of the average Type I error rate for l_z across EA levels for low and mid T levels. The graph shows that there is a steep drop in Type I error rates as EA increases from 10% to 25% for both low and mid T levels and then begin to slowly increase as EA increases to 50%. In general, the Type I error rates are lower for the low T level versus the mid T level.

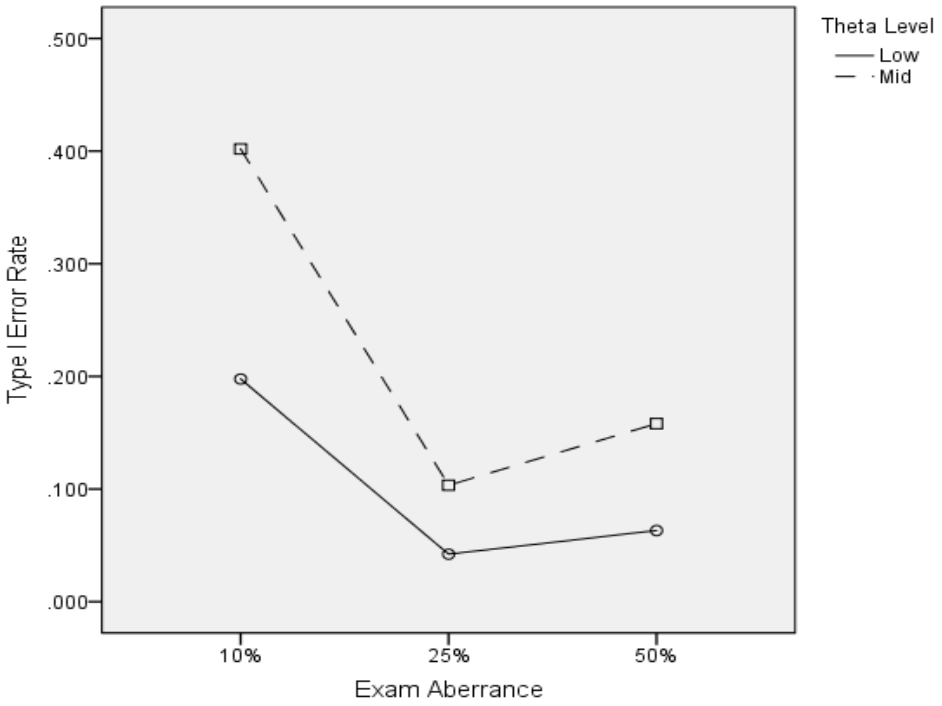

Figure 6. Average Type I error rates for l_z under condition EA x T.

Finally, an analysis of the three-way interaction EL x EA x T with l_z showed statistically significant differences in EA means across the combined EL and T conditions. Specifically, results for the short EL condition, low and mid T levels were $F(2, 23952) = 26102.561$, $p < .001$ and $F(2, 23952) = 30587.745$, $p < .001$, respectively.

Results for the long EL condition, low and mid T levels were F(2, 23952) = 1470.370, p < .001 and F(2, 23952) = 44045.178, p < .001, respectively. Simple pairwise comparisons showed statistically significant mean differences at p < .001 between all levels of EA for a given EL x T condition, with the exception of the 25% and 50% EA mean differences for the long EL, low T condition.

Figures 7 and 8 show the Type I error rates for EL x EA for each level of T. Similar to the pattern seen in Figure 5 above, the Type I error rates drop as EA increases for the long EL condition at both T levels. However, for the low T level, this drop is more pronounced for the 10% EA condition whereas for the mid T level, the drop is more pronounced for the 25% and 50% EA conditions.

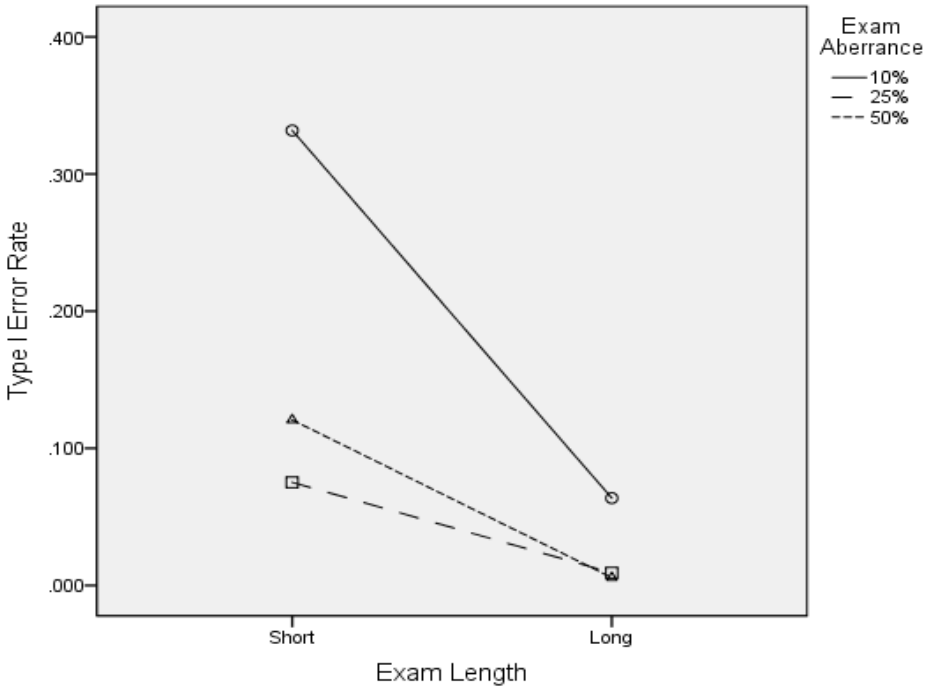

Figure 7. Average Type I error rates for l_z under condition EL x EA x T for T = Low.

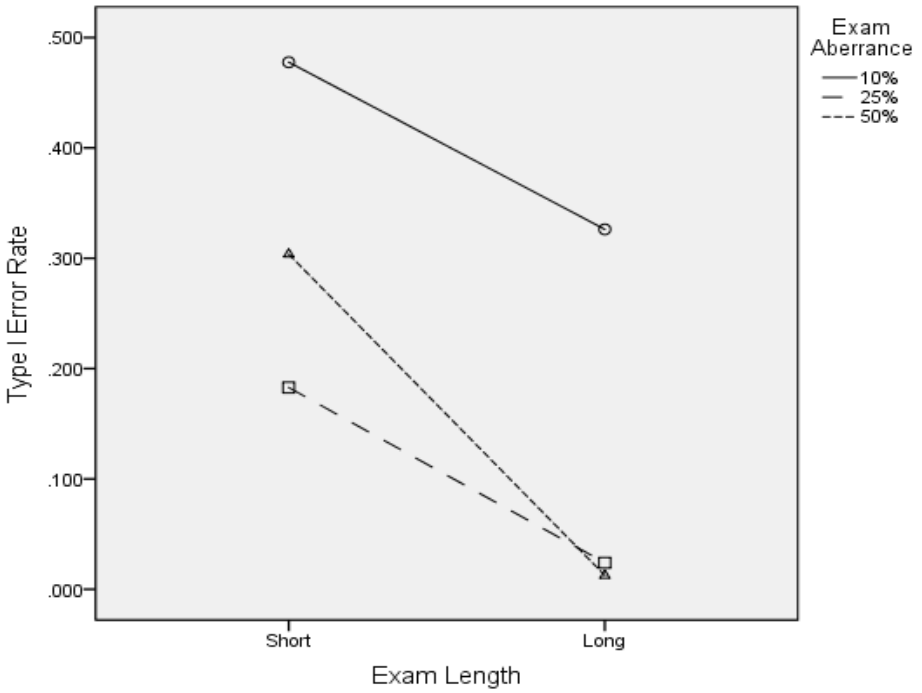

Figure 8. Average Type I error rates for l_z under condition EL x EA x T for T = Mid.

There was one two-way interaction for the ERT measure that yielded a small effect size, namely EA x T. A simple effects analysis for EA x T with ERT showed statistically significant differences between EA means for both the low ($F(2, 23952) = 195131.773$, $p < .001$) and mid ($F(2, 23952) = 94299.462$, $p < .001$) T levels. Simple pairwise comparisons showed statistically significant mean differences at $p < .001$ between all levels of EA within both the low and mid T levels. Figure 9 below shows the Type I error rates for ERT under the EA x T condition. As with l_z, the Type I error rates for ERT across both low and mid T levels drop as EA increases from 10% to 25%, and then begin to increase as EA increases to 50%. While the changes in Type I error rates for ERT appear to be more pronounced than those for l_z, they are also notably higher for the 10% and 50% EA conditions.

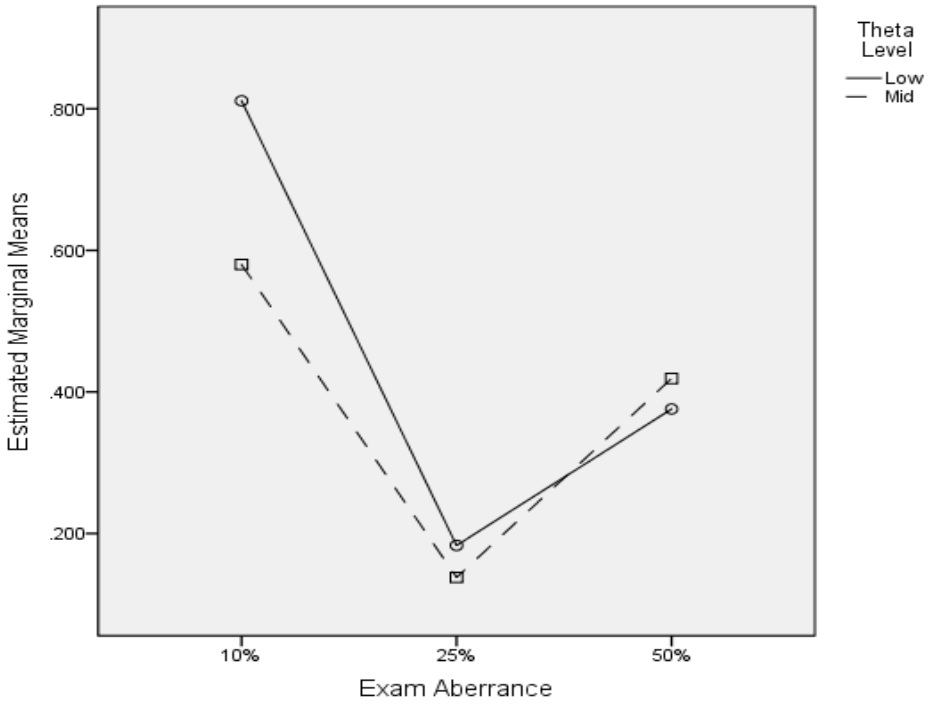

Figure 9. Average Type I error rates for ERT under condition EA x T.

Type I Error Rate for Multiple Measures

Tables containing the full set of results for the factorial ANOVA evaluating the impact of study factors on Type I error rates for the multiple person-fit measures are presented in Appendix B. The sum of squares (SS) and estimated effect sizes (*f*) for interaction effects showing at least a small effect size for multiple person-fit measures are provided in Table 16.

Table 16: *Sums of Squares (SS) and Effect Sizes (f) for Multiple Measures Yielding a Small to Medium f for Interaction Effects*

	l_z + ERT		l_z + PRC		ERT + PRC		l_z + ERT + PRC	
Effect	SS	*f*	SS	*f*	SS	*f*	SS	*f*
M x EA	0.963	0.05	5.902	0.07	30.830	**0.11**	1.170	0.06
EL x EA	23.928	**0.25**	10.488	**0.10**	22.935	**0.10**	19.910	**0.25**
EA x SA	0.200	0.02	0.000	0.00	26.373	**0.11**	0.207	0.02
EA x T	3.274	0.09	25.414	**0.15**	6.846	0.05	4.310	**0.11**
EL x EA x T	6.125	**0.12**	16.274	**0.12**	3.178	0.04	3.967	**0.11**

There were no large effect sizes yielded by any combination of person-fit measures for the interaction effects. l_z + ERT and l_z + ERT + PRC produced medium effect sizes for the EL x EA condition, with f = 0.25 in both cases. Small effect sizes for the EL x EA condition were produced by l_z + PRC and ERT + PRC, with f = .10 for both. The EL x EA x T condition yielded small effect sizes across three sets of combined measures, with f = 0.12 for l_z + ERT and l_z + PRC, and f = 0.11 for l_z + ERT + PRC. Small effect sizes were also produced by l_z + PRC and l_z + ERT + PRC in the EA x T condition, with f = 0.15 and f = 0.11, respectively. Small effect sizes were produced for ERT + PRC under the conditions EA x SA and M x EA, with f = 0.11 in both cases.

Tables 17 – 21 provide the average Type I error rates for the multiple measures yielding small to medium effect sizes under a given interaction.

Table 17: *Multiple Measures Average Type I Error Rate for EL x EA Interaction*

EL	EA	l_z + ERT	l_z + PRC	ERT + PRC	l_z + ERT + PRC
Short	10%	0.265	0.370	0.473	0.238
	25%	0.024	0.128	0.128	0.024
	50%	0.101	0.212	0.434	0.101
Long	10%	0.088	0.174	0.221	0.074
	25%	0.001	0.016	0.027	0.001
	50%	0.003	0.009	0.255	0.003

Note: Standard error was 0.000 for l_z+ERT and l_z+ERT+PRC average Type I Error rates. Standard error was 0.001 for l_z+PRC and ERT+PRC average Type I Error rates. EL = exam length. EA = exam aberrance.

Table 18: *Multiple Measures Average Type I Error Rate for EL x EA x T Interaction*

EL	EA	T	l_z + ERT	l_z + PRC	l_z + ERT + PRC
Short	10%	Low	0.246	0.282	0.206
		Mid	0.285	0.459	0.271
	25%	Low	0.012	0.073	0.012
		Mid	0.036	0.182	0.035
	50%	Low	0.050	0.121	0.050
		Mid	0.153	0.304	0.153
Long	10%	Low	0.040	0.048	0.030
		Mid	0.135	0.299	0.119
	25%	Low	0.000	0.009	0.000
		Mid	0.001	0.024	0.001
	50%	Low	0.001	0.006	0.001
		Mid	0.004	0.013	0.004

Note: Standard error was 0.001 for all average Type I Error rates.
EL = exam length. EA = exam aberrance. T = theta level.

Table 19: *Multiple Measures Average Type I Error Rate for EA x T Interaction*

EA	T	l_z + *PRC*	l_z + *ERT* + *PRC*
10%	Low	0.165	0.118
	Mid	0.379	0.195
25%	Low	0.041	0.006
	Mid	0.103	0.018
50%	Low	0.063	0.025
	Mid	0.158	0.078

Note: Standard error was 0.001 for l_z+PRC average Type I Error rates. Standard error was 0.000 for l_z+ERT+PRC average Type I Error rates.
EA = exam aberrance. T = theta level.

Table 20: *ERT + PRC Average Type I Error Rate for EA x SA Interaction*

EA	SA	*ERT + PRC*
10%	5%	0.335
	10%	0.359
25%	5%	0.053
	10%	0.103
50%	5%	0.256
	10%	0.433

Note: Standard error was 0.001 for all average Type I Error rates.
EA = exam aberrance. SA = sample aberrance.

Table 21: *ERT + PRC Average Type I Error Rate for M x EA Interaction*

Model	EA	*ERT + PRC*
Rasch	10%	0.291
	25%	0.081
	50%	0.374
2PL	10%	0.403
	25%	0.075
	50%	0.315

Note: Standard error was 0.001 for all average Type I Error rates.
EA = exam aberrance.

Figures 10 – 23 below provide graphical representation of the average Type I error rate for the multiple person-fit measures showing small to medium interaction effect sizes. As with the individual measures, simple effects analysis was performed to further explore the two-way and three-way interaction effects for multiple measure Type I error rates.

The results of the analysis for EL x EA with l_z + ERT showed statistically significant differences in EA means for both the short ($F(2, 23952) = 76478.453$, $p < .001$) and long ($F(2, 23952) = 12374.692$, $p < .001$) EL conditions. Simple pairwise comparisons further revealed statistically significant mean differences at $p < .001$

between all levels of EA for the short EL condition and between the 10% and 25% and the 10% and 50% levels of EA for the long EL condition. The results of the analysis for EL x EA with l_z + PRC also showed statistically significant differences in EA means for both the short ($F(2, 23952) = 43157.005$, $p < .001$) and long ($F(2, 23952) = 24606.013$, $p < .001$) EL conditions. Simple pairwise comparisons showed statistically significant mean differences at $p < .001$ between all levels of EA for both the short and long EL conditions.

Figures 10 and 11 show the Type I error rate for l_z x ERT and l_z x PRC, respectively. Using both l_z and ERT to identify aberrance resulted in lower Type I error rates across all EL x EA conditions. However, as can be seen by comparing Figure 11 to Figure 5, combining l_z and PRC showed little to no difference on Type I error rates than when l_z alone is used.

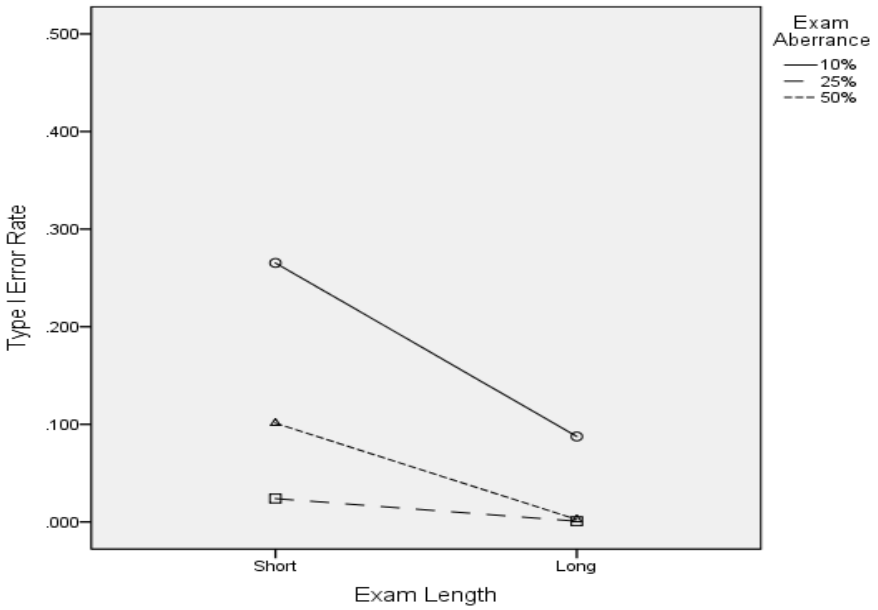

Figure 10. Average Type I error rates for l_z + ERT under condition EL x EA.

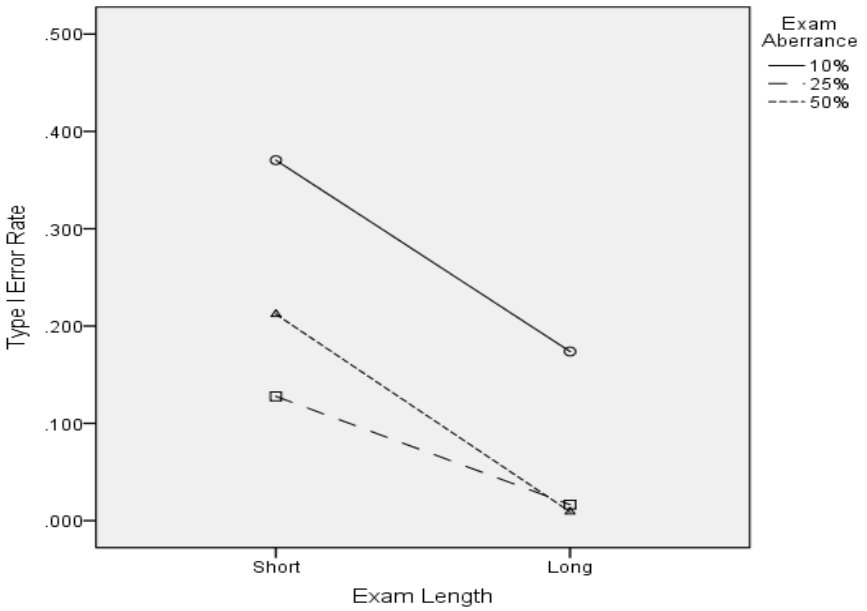

Figure 11. Average Type I error rates for l_z + PRC under condition EL x EA.

Simple effects analysis of EL x EA with ERT + PRC showed statistically significant differences in EA means for both the short ($F(2, 23952) = 75200.880$, $p < .001$) and long ($F(2, 23952) = 31858.861$, $p < .001$) EL conditions. In addition, simple pairwise comparisons revealed statistically significant mean differences at $p < .001$ between all levels of EA for both the short and long EL conditions. Figure 12 shows an interesting interaction between EL and the 10% and 50% EA conditions on Type I error rates when using ERT + PRC to identify aberrance. Specifically, the Type I error rate is slightly higher for the short EL x 10% EA condition than it is for the short EL x 50% EA condition, but is slightly lower under the long EL condition.

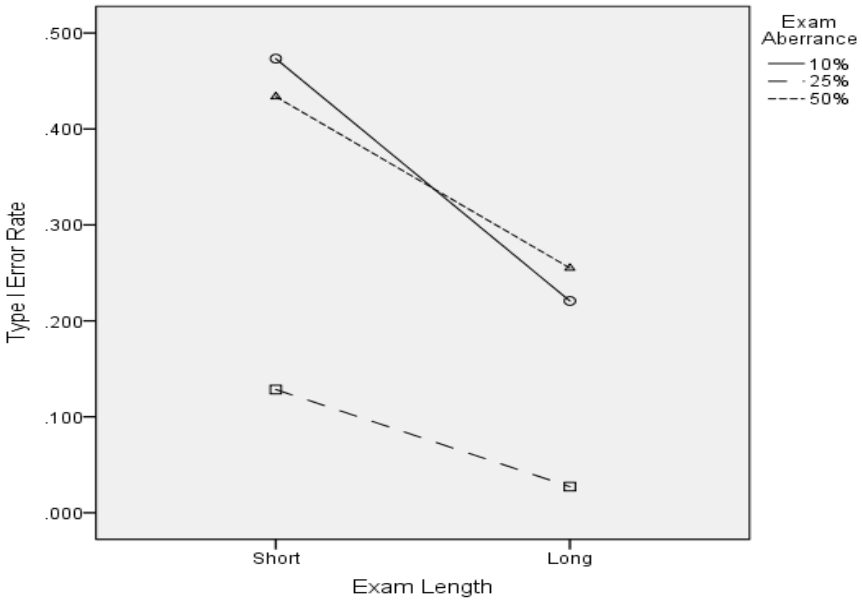

Figure 12. Average Type I error rates for ERT + PRC under condition EL x EA.

Finally, using PRC in addition to l_z and ERT to detect aberrance did not appear to have a noticeable impact on Type I error rates under the EL x EA condition. This can be seen by comparing Figure 13 below to Figure 10. Results of the simple effects analysis yielded statistically significant differences in EA means for both the short ($F(2, 23952) = 61744.998$, $p < .001$) and long ($F(2, 23952) = 9209.005$, $p < .001$) EL conditions. In addition, the results of the simple pairwise comparisons of mean differences were statistically significant at $p < .001$ between all levels of EA for the short EL condition. However, for the long EL condition, mean differences were statistically significant only for two conditions – the 10% and 25% and 10% and 50% levels of EA for the long EL condition.

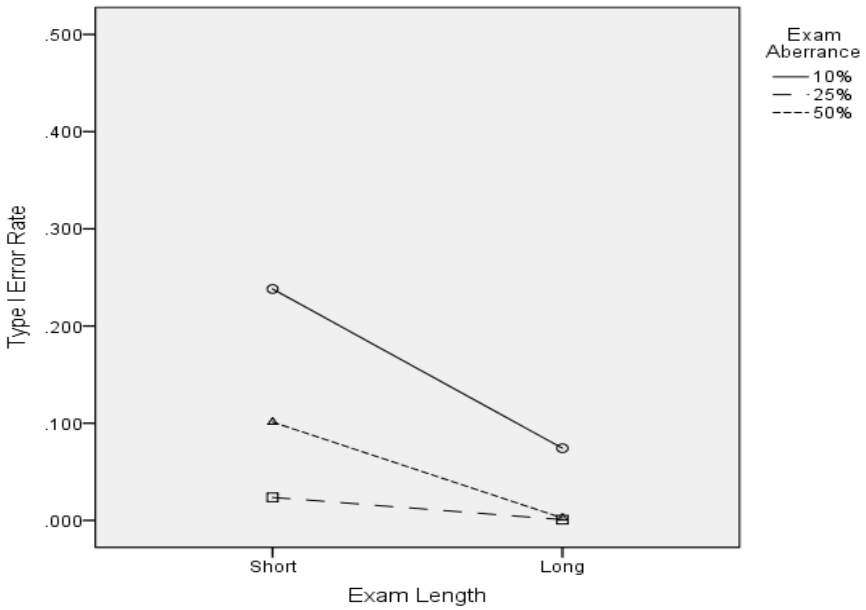

Figure 13. Average Type I error rates for l_z + ERT + PRC under condition EL x EA.

The patterns in Type I error rates for l_z + ERT and l_z + ERT + PRC under the EL x EA x T condition are similar to those found for the same combinations of measures under the EL x EA condition. It should be noted that for both l_z + ERT and l_z + ERT + PRC, there is a more pronounced decline in Type I error rates going from a short to long exam length for the 25% and 50% EA, mid level T conditions. Figures 14 – 17 show this pattern for the l_z + ERT and l_z + ERT + PRC by T level, respectively.

The simple effects analysis for the three-way interaction EL x EA x T with l_z + ERT showed statistically significant differences in EA means across combined EL and T conditions. Results for the short EL condition, low and mid T levels were $F(2, 23952) = 39563.429$, $p < .001$ and $F(2, 23952) = 39157.364$, $p < .001$, respectively. Results for the long EL condition, low and mid T levels were $F(2, 23952) = 1282.880$, $p < .001$ and $F(2, 23952) = 14762.753$, $p < .001$, respectively. Simple pairwise comparisons showed statistically significant mean differences at $p < .001$ between all levels of EA within EL

and T, with the exception of the 25% and 50% EA mean differences within the long EL, low and mid T conditions.

Similarly, the simple effects analysis for the three-way interaction EL x EA x T with l_z + ERT + PRC showed statistically significant differences in EA means across combined EL and T conditions. Results for the short EL condition, low and mid T levels were F(2, 23952) = 27607.616, $p < .001$ and F(2, 23952) = 36207.123, $p < .001$, respectively. Results for the long EL condition, low and mid T levels were F(2, 23952) = 725.550, $p < .001$ and F(2, 23952) = 11832.482, $p < .001$, respectively. Finally, the results of the simple pairwise comparisons were the same as those described above for l_z + ERT.

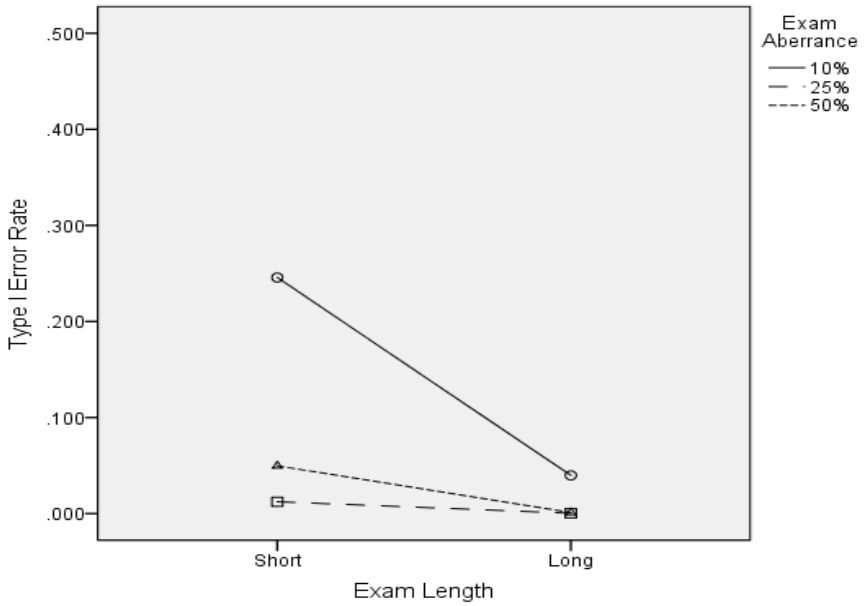

Figure 14. Average Type I error rates for l_z + ERT under condition EL x EA x T for T = Low.

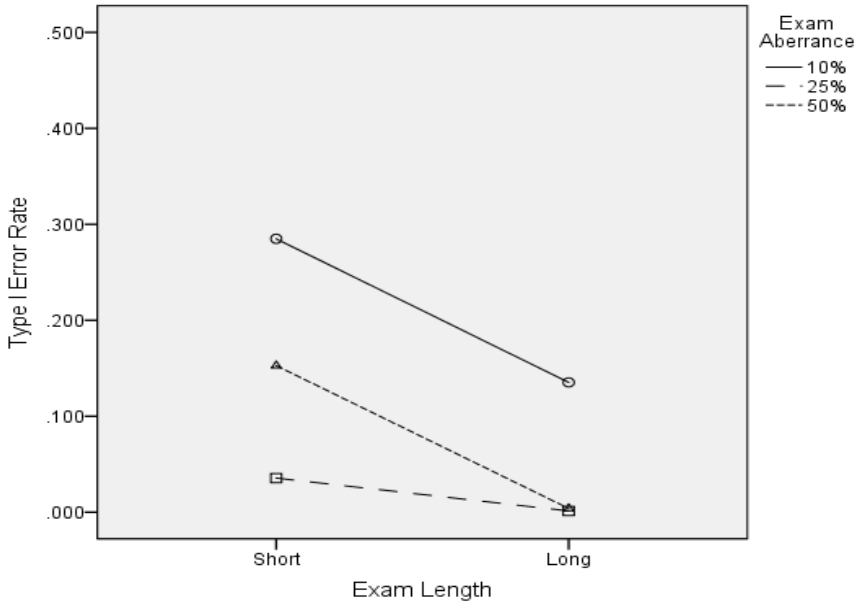

Figure 15. Average Type I error rates for l_z + ERT under condition EL x EA x T for T = Mid.

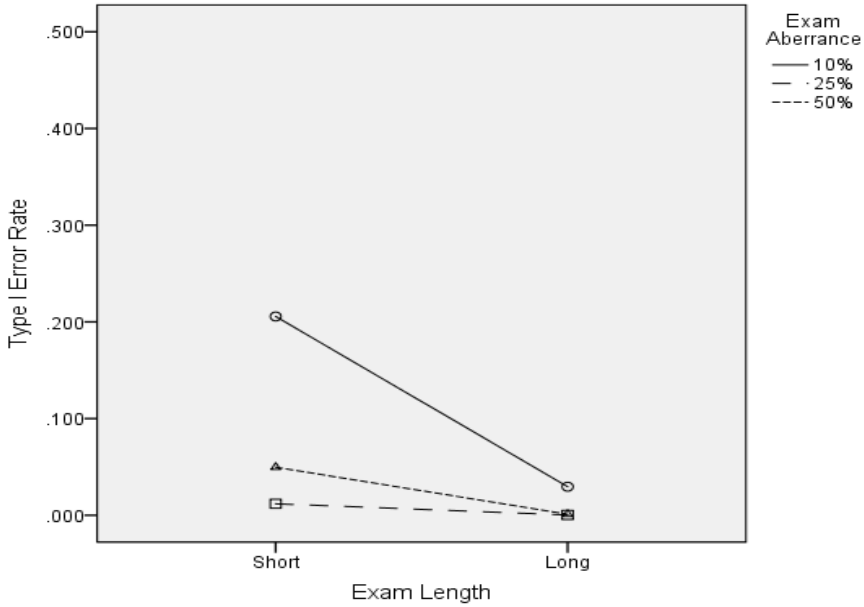

Figure 16. Average Type I error rates for l_z + ERT + PRC under condition EL x EA x T for T = Low.

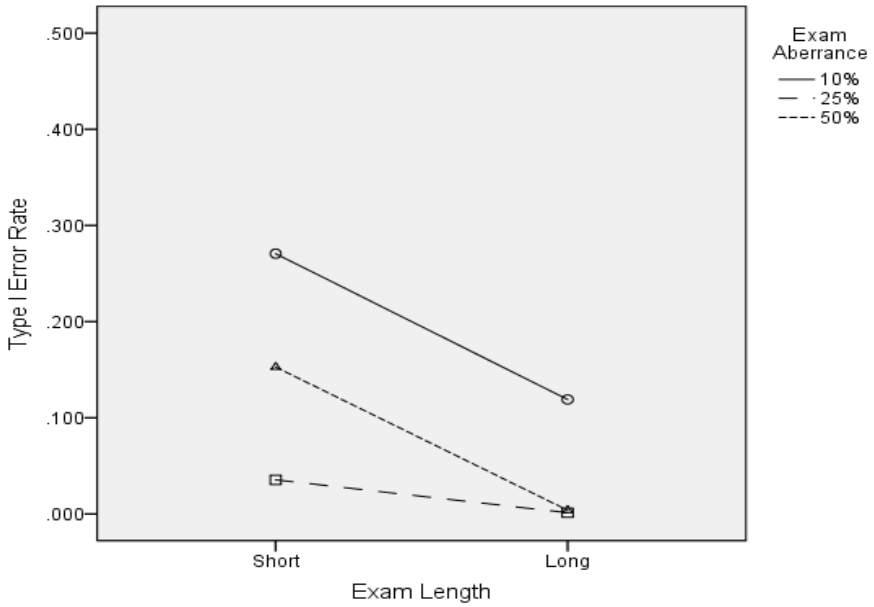

Figure 17. Average Type I error rates for l_z + ERT + PRC under condition EL x EA x T for T = Mid.

Figures 18 and 19 provide graphical representation of the interaction *EL x EA x T* for l_z + PRC. It can be seen that Type I error rates were lower overall for l_z + PRC under the low T condition (across EL and EA conditions). However, there is a more pronounced decline in Type I error rates for the long EL, mid T level under the 25% and 50% EA condition. Simple effects analysis for l_z + PRC showed statistically significant differences with $F(2, 23952) = 17058.090$, $p < .001$ and $F(2, 23952) = 27293.168$, $p < .001$ for the short EL condition, low and mid T levels, respectively, and $F(2, 23952) = 806.689$, $p < .001$ and $F(2, 23952) = 37422.479$, $p < .001$ for the long EL condition, low and mid T levels, respectively.

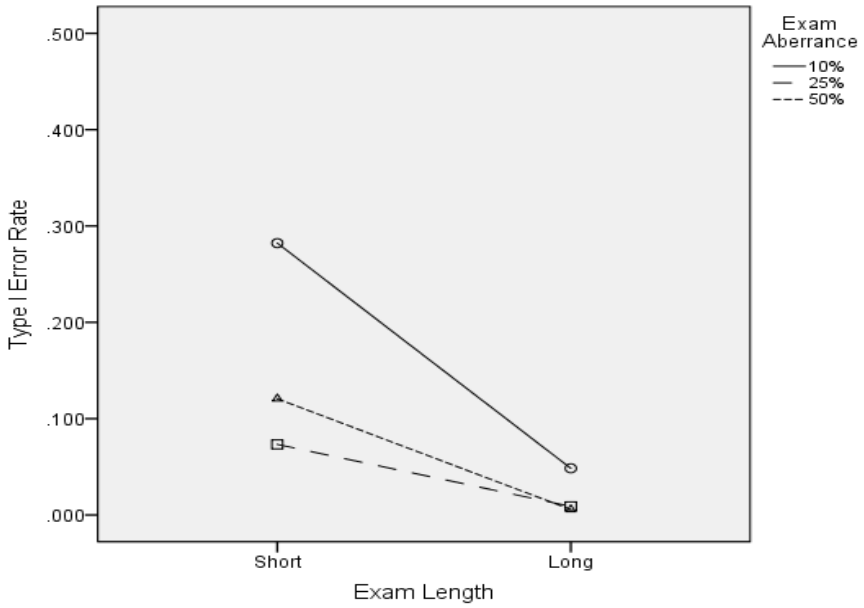

Figure 18. Average Type I error rates for l_z + PRC under condition EL x EA x T for T = Low.

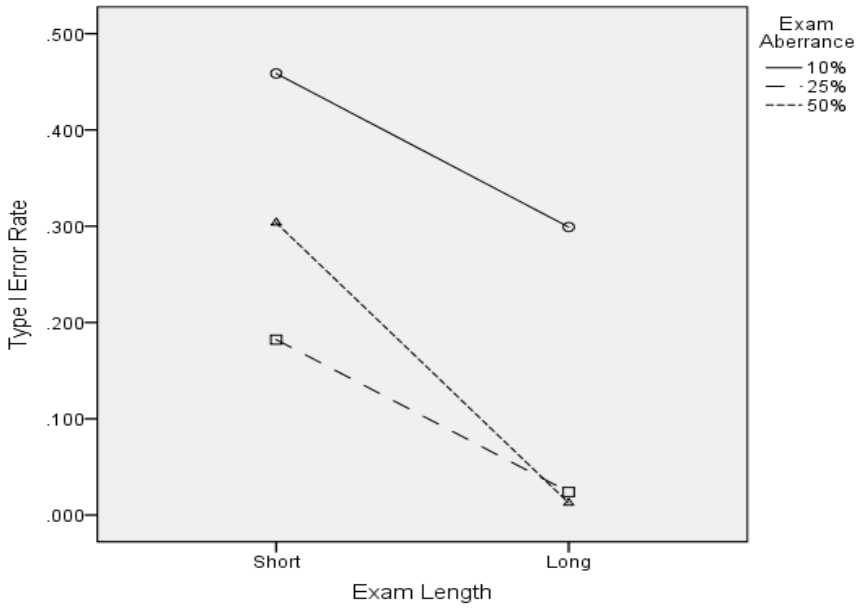

Figure 19. Average Type I error rates for l_z + PRC under condition EL x EA x T for T = Mid.

In reviewing the graph for l_z + PRC under the EA x T condition shown in Figure 20, it is apparent that the pattern is very close to that shown for l_z alone under the EA x T

condition in Figure 6. It appears that adding PRC as a flag for aberrance does not impact Type I error rates of the interaction effects under this condition. The results of the EA simple effects analysis were statistically significant for both T levels, with F(2, 23952) = 12491.304, $p < .001$ for the low T level and F(2, 23952) = 60576.686, $p < .001$ for the mid T level. Simple pairwise comparison revealed that all mean differences were statistically significant at $p < .001$.

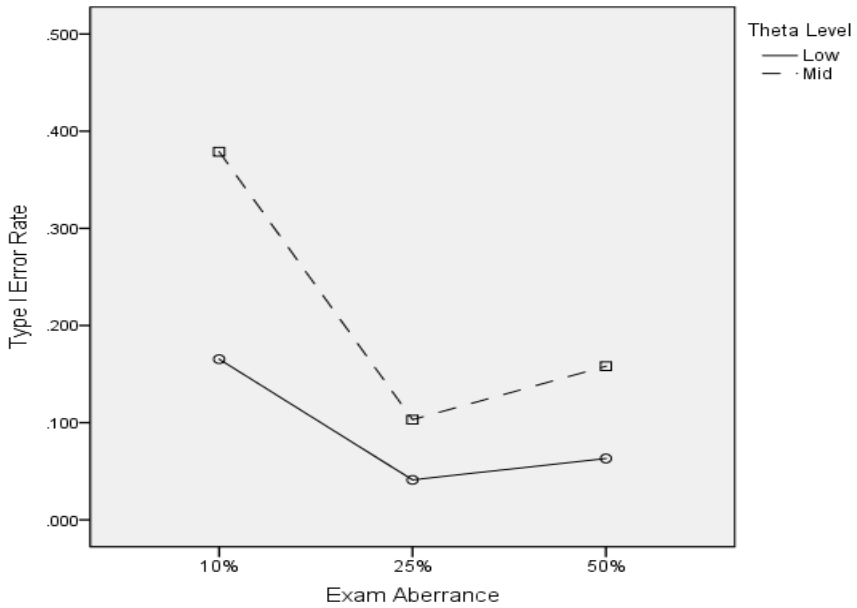

Figure 20. Average Type I error rates for l_z + PRC under condition EA x T.

In reviewing Type I error rates for l_z + ERT + PRC under the EA x T condition (Figure 21 below), it can be seen that the error rates are lower under the low T level compared to the mid T level, with the 25% EA condition producing the lowest Type I error rates (as compared to the 10% and 50% EA levels). The results of the EA simple effects analysis with l_z + ERT + PRC were statistically significant for both T levels, with F(2, 23952) = 18584.461, $p < .001$ for the low T level and F(2, 23952) = 42156.084, $p <$

.001 for the mid T level. In addition, all simple pairwise comparisons of the mean differences were statistically significant at $p < .001$.

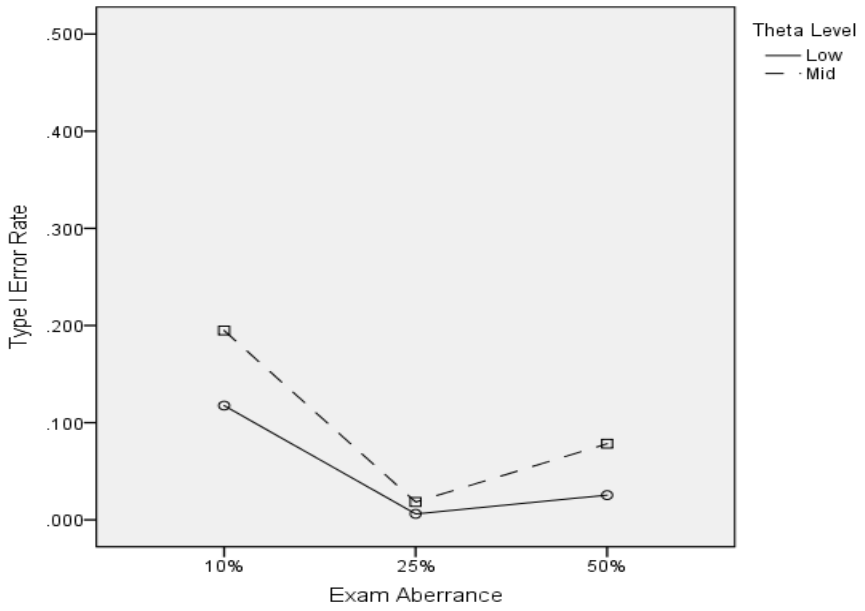

Figure 21. Average Type I error rates for l_z + ERT + PRC under condition EA x T.

While there were no significant interaction effects for ERT or PRC under the EA x SA or the M x EA conditions, using these two measures together resulted in small effect sizes under both conditions. Figures 22 and 23 show the Type I error rates for ERT + PRC under the EA x SA and M x EA conditions, respectively.

Under the EA x SA condition, ERT + PRC produced lower error rates under the 5% SA level as compared to the 15% level, with the 25% EA level producing the lowest Type I error rates (as compared to the 10% and 50% EA levels). The EA simple effects analysis resulted in $F(2, 23952) = 44704.468$, $p < .001$ for 5% SA and $F(2, 23952) = 63262.772$, $p < .001$ for 15% SA, with pairwise comparisons producing statistically significant results at $p < .001$ across all EA x SA conditions.

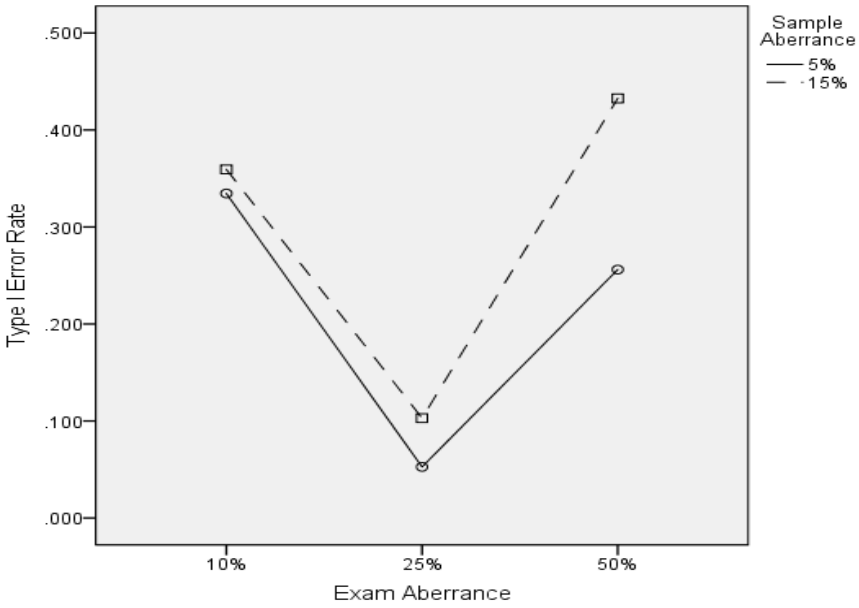

Figure 22. Average Type I error rates for ERT + PRC under condition EA x SA.

Finally, ERT + PRC produced the lowest Type I error rates under the 25% EA level, however there appeared to be no real difference between the Rasch and 2PL IRT models. There was a difference in IRT models for the 10% and 50% EA levels, with the error rate increasing for the 2PL model under the 10% EA level and decreasing under the 50% EA level. The EA simple effects analysis showed statistically significant results with $F(2, 23952) = 48084.993$, $p < .001$ for the Rasch model and $F(2, 23952) = 61058.456$, $p < .001$ for the 2PL model and simple pairwise comparisons showed statistically significant mean differences ($p < .001$) across all M x EA conditions.

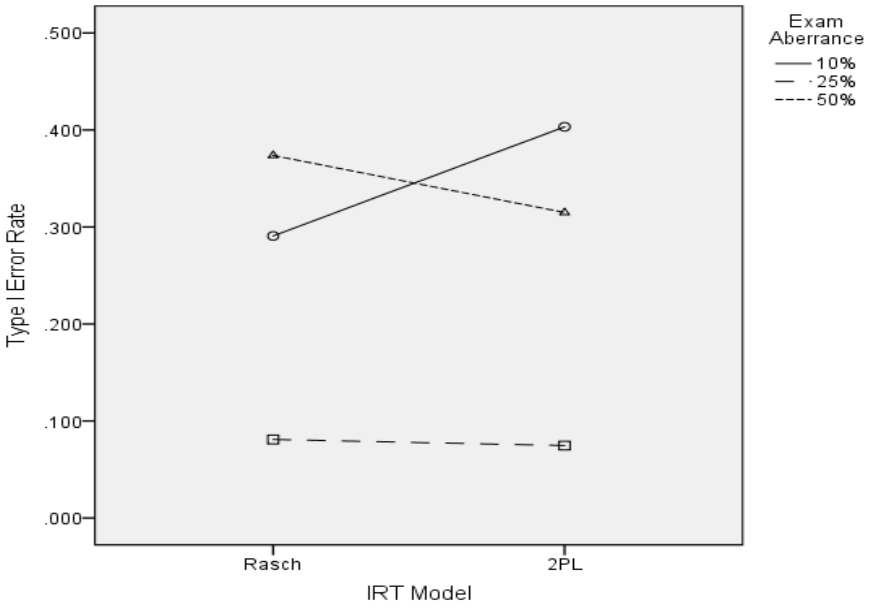

Figure 23. Average Type I error rates for ERT + PRC under condition M x EA.

Classification Accuracy

Coefficient κ was computed as a measure of the proportion of classification agreement between individual person-fit measures. The equations used to compute κ and the standard error of κ are provided in Equations 15 and 16, respectively. Tables 22, 24, and 26 below provide the mean value of κ for each combination of person-fit measures for each study condition and Tables 23, 25, 27 provide the corresponding standard deviation of the mean. Figures 24-25 below provide graphs showing the mean value of κ for each combination of person-fit measures at each level of EA (10%, 25%, 50%) by T level (low, mid) within each study condition SA x M x EL.

The graphs and information provided in the tables below show that l_z and PRC had the highest average values for κ (compared to l_z and ERT or ERT and PRC) across study conditions when EA was set to 10% or 25% and T level was low. l_z and ERT also exhibited strong κ values across conditions when EA was set to 25% and T level was

low, followed by 25% EA and mid T levels. ERT and PRC exhibited very low κ values overall, with negative values appearing when EA was set to 50% and T was at mid level.

Table 22: *Mean kappa: l_z and ERT*

			Theta level: Low			Theta level: Mid		
			Exam aberrance			Exam aberrance		
Model	EL	SA	10%	25%	50%	10%	25%	50%
Rasch	Short	5%	0.032	0.317	0.247	0.071	0.318	0.177
		15%	0.061	0.492	0.316	0.115	0.464	0.210
	Long	5%	0.083	0.469	0.429	0.152	0.452	0.388
		15%	0.133	0.737	0.589	0.223	0.668	0.459
2PL	Short	5%	0.039	0.335	0.293	0.085	0.358	0.240
		15%	0.067	0.516	0.360	0.134	0.510	0.264
	Long	5%	0.077	0.475	0.433	0.175	0.467	0.405
		15%	0.132	0.749	0.588	0.265	0.697	0.488

Note: EL = exam length. SA = sample aberrance.

Table 23: *Standard Deviation of the Mean kappa: l_z and ERT*

			Theta level: Low			Theta level: Mid		
			Exam aberrance			Exam aberrance		
Model	EL	SA	10%	25%	50%	10%	25%	50%
Rasch	Short	5%	0.036	0.038	0.042	0.040	0.041	0.043
		15%	0.035	0.034	0.037	0.042	0.038	0.042
	Long	5%	0.038	0.033	0.037	0.042	0.032	0.039
		15%	0.036	0.026	0.032	0.039	0.028	0.033
2PL	Short	5%	0.038	0.043	0.044	0.042	0.040	0.047
		15%	0.037	0.036	0.037	0.043	0.037	0.039
	Long	5%	0.038	0.032	0.037	0.044	0.031	0.036
		15%	0.035	0.025	0.030	0.039	0.025	0.034

Note: EL = exam length. SA = sample aberrance.

Table 24: *Mean kappa: l_z and PRC*

			Theta level: Low			Theta level: Mid		
			Exam aberrance			Exam aberrance		
Model	EL	SA	10%	25%	50%	10%	25%	50%
Rasch	Short	5%	0.445	0.461	0.235	0.328	0.290	0.175
		15%	0.570	0.567	0.211	0.391	0.312	0.097
	Long	5%	0.508	0.481	0.226	0.367	0.276	0.183
		15%	0.715	0.664	0.143	0.440	0.240	0.040
2PL	Short	5%	0.382	0.470	0.321	0.279	0.288	0.186
		15%	0.465	0.576	0.349	0.299	0.298	0.124
	Long	5%	0.497	0.542	0.379	0.325	0.323	0.192
		15%	0.645	0.714	0.435	0.333	0.318	0.075

Note: EL = exam length. SA = sample aberrance.

Table 25: *Standard Deviation of the Mean kappa: l_z and PRC*

			Theta level: Low			Theta level: Mid		
			Exam aberrance			Exam aberrance		
Model	EL	SA	10%	25%	50%	10%	25%	50%
Rasch	Short	5%	0.053	0.049	0.053	0.058	0.056	0.052
		15%	0.037	0.034	0.037	0.047	0.041	0.034
	Long	5%	0.035	0.038	0.043	0.041	0.043	0.040
		15%	0.026	0.030	0.036	0.036	0.036	0.030
2PL	Short	5%	0.057	0.052	0.055	0.057	0.054	0.051
		15%	0.044	0.035	0.042	0.047	0.040	0.035
	Long	5%	0.044	0.039	0.048	0.050	0.048	0.043
		15%	0.034	0.027	0.038	0.043	0.036	0.031

Note: EL = exam length. SA = sample aberrance.

Table 26: *Mean kappa: ERT and PRC*

			Theta level: Low			Theta level: Mid		
			Exam aberrance			Exam aberrance		
Model	EL	SA	10%	25%	50%	10%	25%	50%
Rasch	Short	5%	0.030	0.214	0.042	0.024	0.094	-0.002
		15%	0.053	0.339	0.059	0.033	0.143	-0.004
	Long	5%	0.061	0.279	0.013	0.049	0.061	-0.021
		15%	0.106	0.536	0.013	0.065	0.093	-0.053
2PL	Short	5%	0.023	0.222	0.106	0.022	0.101	0.015
		15%	0.038	0.347	0.148	0.031	0.148	0.018
	Long	5%	0.055	0.364	0.177	0.040	0.128	-0.007
		15%	0.093	0.622	0.285	0.050	0.204	-0.026

Note: EL = exam length. SA = sample aberrance.

Table 27: *Standard Deviation of the Mean kappa: ERT and PRC*

			Theta level: Low			Theta level: Mid		
			Exam aberrance			Exam aberrance		
Model	EL	SA	10%	25%	50%	10%	25%	50%
Rasch	Short	5%	0.036	0.046	0.037	0.034	0.040	0.028
		15%	0.038	0.043	0.036	0.035	0.039	0.026
	Long	5%	0.033	0.034	0.032	0.036	0.035	0.028
		15%	0.032	0.031	0.032	0.035	0.033	0.025
2PL	Short	5%	0.034	0.050	0.045	0.034	0.039	0.030
		15%	0.035	0.040	0.047	0.036	0.039	0.030
	Long	5%	0.036	0.038	0.045	0.037	0.040	0.030
		15%	0.037	0.030	0.042	0.037	0.036	0.026

Note: EL = exam length. SA = sample aberrance.

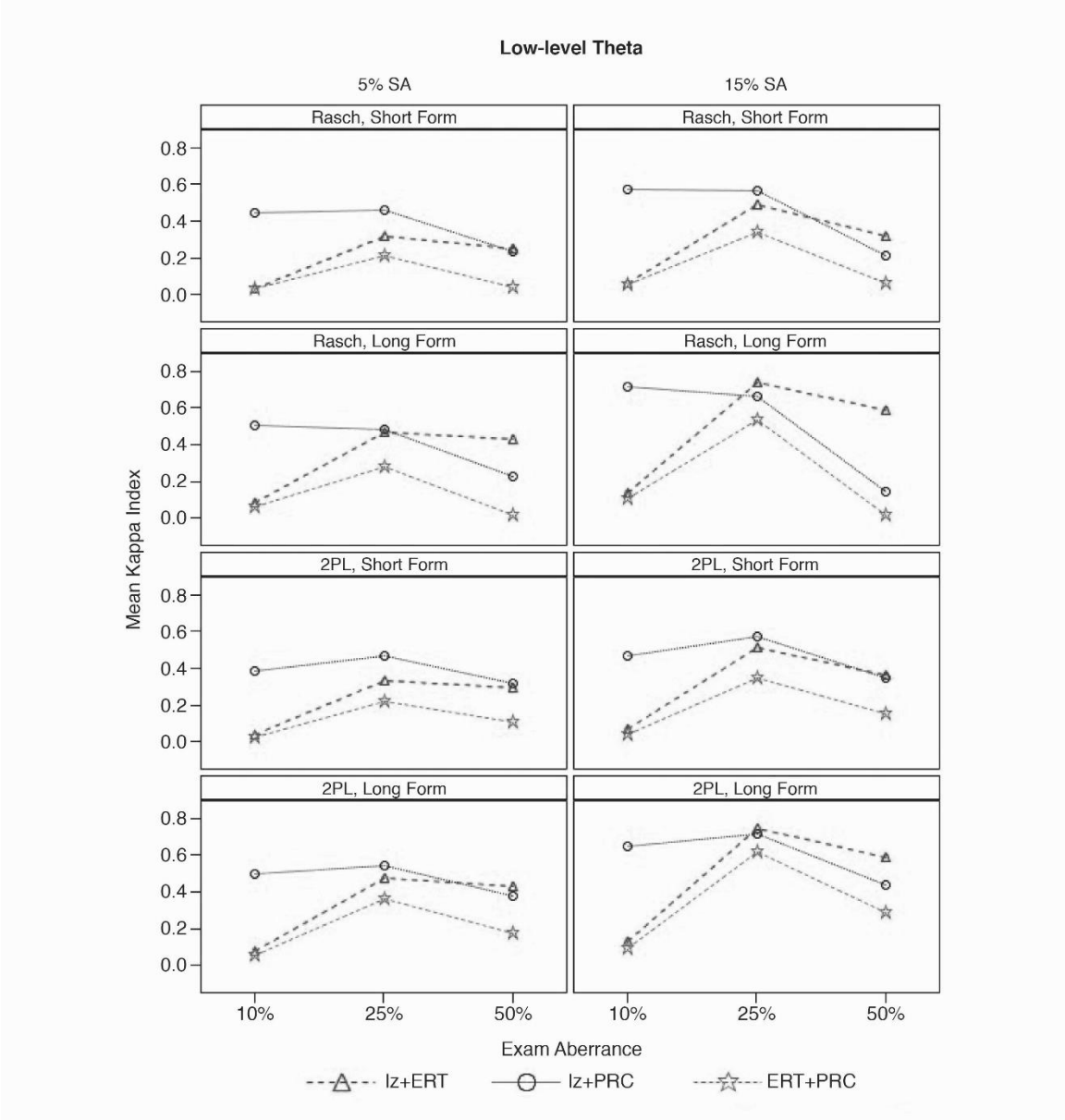

Figure 24. Mean kappa values for low-level theta condition across study factors.

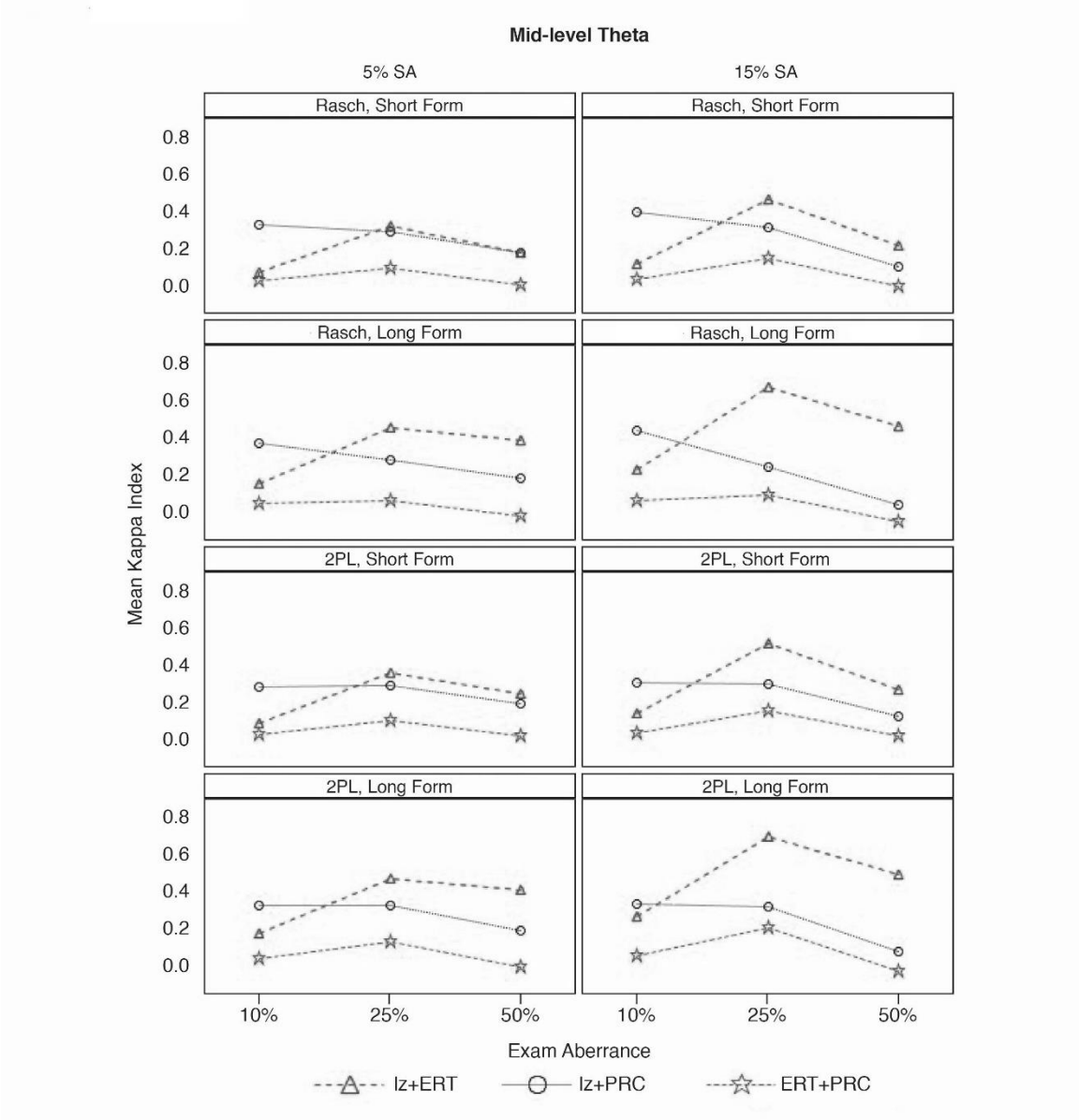

Figure 25. Mean kappa values for mid-level theta condition across study factors.

Sensitivity

Sensitivity was computed for individual and combined person-fit measures. Mean sensitivity values for individual person-fit measures are provided in Tables 28 and 30 for low- and mid-level T conditions, respectively, and Tables 29 and 31 contain corresponding values for the standard deviation of the mean. Sensitivity values for multiple-measure combinations are in Tables 32 and 34 for low- and mid-level T conditions, respectively, with corresponding standard deviation values in Tables 33 and 35. Appendix C contains figures providing graphical representation of the mean sensitivity values for each study condition.

As an individual measure, l_z provided high sensitivity values for all low-level T, long EL conditions, with values ranging from 0.935 to 1.000. The low-level T, short EL conditions in which l_z performed well included the 25% EA conditions (values ranged from 0.902 to 0.948) and the 50% EA condition for the 2PL model only (value = 0.941). Under the mid-level T condition, l_z performed well at the 25% and 50% EA, long EL conditions, with values ranging from 0.961 to 0.991.

ERT exhibited good sensitivity under limited conditions. Specifically, values for ERT ranged from 0.891 to 0.972 under low- and mid-level T, 25% EA, long EL conditions. Similarly, PRC performed well under limited conditions, namely the low-level T, 10% EA, long EL conditions for the Rasch model only. The values for PRC under this condition were 0.907 for 5% SA and 0.905 for 15% SA. PRC also had notably the lowest sensitivity values under mid-level T, 50% EA conditions, ranging from 0.001 to 0.052.

The combined (i.e., multiple) measures performed best under the low-level T, 25% EA conditions, with values ranging from 0.882 to 1.000. Three sets of combined measures (l_z + ERT, l_z + PRC, l_z + ERT + PRC) also performed well under the low-level T, long EL conditions for the 10% and 50% EA conditions, with sensitivity values ranging

from 0.943 to 1.000. ERT + PRC also had high sensitivity values under the low level T, long EL, Rasch model conditions. The values for ERT + PRC under these conditions were 0.942 for 5% SA and 0.938 for 15% SA. Sensitivity values under the low level T, short EL conditions were also high for l_z + ERT, l_z + PRC, and l_z + ERT + PRC under the 50% EA conditions, ranging from 0.818 to 0.981.

The lowest sensitivity values for the mid level T conditions occurred under the 10% EA conditions across all combined measures, with values ranging from 0.697 to 0.886 for l_z + ERT, 0.528 to 0.704 for l_z + PRC, 0.454 to 0.778 for ERT + PRC, and 0.716 to 0.892 for l_z + ERT + PRC. Conversely, l_z + ERT had notably high levels of sensitivity (0.942 to 1.000) for the majority of the mid level T, 25% and 50% EA conditions. Sensitivity values were slightly lower for l_z + ERT under mid level T, 50% EA, short EL conditions for the Rasch model (sensitivity = 0.841 for 5% SA and 0.761 for 15% SA) and under the 15% SA condition for the 2PL model (sensitivity = 0.870).

It was noted that sensitivity values for l_z + ERT + PRC were very close, if not the same, to those for l_z + ERT across the 25% and 50% EA conditions. This indicates that there was no incremental value in adding PRC to l_z + ERT. In addition, PRC most likely did not contribute to the high sensitivity values produced by l_z + PRC.

Specificity

Specificity values were strongest across all conditions for individual person-fit measures, with values between 0.946 and 0.960 for l_z, 0.890 to 0.964 for ERT, and 0.849 to 0.973 for PRC.

Specificity values for combined person-fit measures were strong, although slightly lower than those for individual measures. Specificity values for ranged from 0.881 to 0.937 for l_z + ERT, 0.833 to 0.933 for l_z + PRC, 0.802 to 0.938 for ERT + PRC, and 0.787 to 0.899 for l_z + ERT + PRC.

Mean specificity values for individual person-fit measures are provided in Tables 36 and 38 for low and mid level T conditions, respectively, and Tables 37 and 39 contain corresponding values for the standard deviation of the mean. Specificity values for multiple-measure combinations are in Tables 40 and 42 for low and mid level T conditions, respectively, with corresponding standard deviation values in Tables 41 and 43. Appendix C contains figures providing graphical representation of the mean specificity values for each study condition.

Table 28: *Sensitivity – Mean Values for Individual Person-Fit Measures Under Low-Level T Conditions*

			Theta level: Low								
			Exam aberrance								
			10%			25%			50%		
Model	EL	SA	l_z	ERT	PRC	l_z	ERT	PRC	l_z	ERT	PRC
Rasch	Short	5%	0.623	0.198	0.505	0.902	0.757	0.532	0.818	0.618	0.126
		15%	0.623	0.185	0.503	0.902	0.687	0.532	0.818	0.435	0.123
	Long	5%	0.937	0.206	0.907	0.985	0.949	0.811	0.989	0.812	0.103
		15%	0.937	0.191	0.905	0.985	0.912	0.812	0.988	0.647	0.102
2PL	Short	5%	0.712	0.179	0.359	0.948	0.721	0.559	0.941	0.608	0.285
		15%	0.715	0.171	0.359	0.948	0.662	0.559	0.941	0.427	0.284
	Long	5%	0.936	0.198	0.679	0.997	0.939	0.811	1.000	0.808	0.415
		15%	0.935	0.183	0.679	0.997	0.907	0.813	1.000	0.638	0.415

Note: EL = exam length. SA = sample aberrance.

Table 29: *Sensitivity – Standard Deviation of the Mean for Individual Person-Fit Measures Under Low-Level T Conditions*

			Theta level: Low								
			Exam aberrance								
			10%			25%			50%		
Model	EL	SA	l_z	ERT	PRC	l_z	ERT	PRC	l_z	ERT	PRC
Rasch	Short	5%	0.070	0.057	0.072	0.042	0.058	0.070	0.055	0.067	0.049
		15%	0.040	0.032	0.041	0.025	0.034	0.040	0.032	0.033	0.027
	Long	5%	0.034	0.059	0.039	0.017	0.032	0.055	0.015	0.051	0.040
		15%	0.019	0.033	0.023	0.010	0.021	0.031	0.009	0.033	0.024
2PL	Short	5%	0.063	0.055	0.068	0.032	0.063	0.073	0.032	0.067	0.065
		15%	0.036	0.031	0.039	0.018	0.036	0.039	0.020	0.035	0.037
	Long	5%	0.033	0.057	0.066	0.008	0.035	0.058	0.002	0.051	0.065
		15%	0.019	0.032	0.038	0.005	0.022	0.031	0.001	0.032	0.039

Note: EL = exam length. SA = sample aberrance.

Table 30: *Sensitivity – Mean Values for Individual Person-Fit Measures Under Mid-Level T Conditions*

			Mid-level Theta								
			Exam aberrance								
			10%			25%			50%		
Model	EL	SA	l_z	ERT	PRC	l_z	ERT	PRC	l_z	ERT	PRC
Rasch	Short	5%	0.530	0.355	0.229	0.784	0.862	0.203	0.610	0.584	0.015
		15%	0.529	0.309	0.229	0.785	0.716	0.207	0.608	0.375	0.015
	Long	5%	0.658	0.499	0.437	0.961	0.966	0.237	0.987	0.774	0.001
		15%	0.660	0.435	0.439	0.961	0.891	0.238	0.987	0.546	0.001
2PL	Short	5%	0.517	0.390	0.144	0.851	0.858	0.211	0.782	0.594	0.052
		15%	0.514	0.344	0.144	0.849	0.720	0.209	0.786	0.371	0.049
	Long	5%	0.686	0.545	0.228	0.991	0.972	0.294	0.987	0.814	0.019
		15%	0.692	0.483	0.230	0.991	0.911	0.296	0.988	0.589	0.019

Note: EL = exam length. SA = sample aberrance.

Table 31: *Sensitivity – Standard Deviation of the Mean for Individual Person-Fit Measures Under Mid-Level T Conditions*

			Theta level: Mid								
			Exam aberrance								
			10%			25%			50%		
Model	EL	SA	l_z	ERT	PRC	l_z	ERT	PRC	l_z	ERT	PRC
Rasch	Short	5%	0.070	0.068	0.059	0.057	0.049	0.056	0.070	0.069	0.018
		15%	0.040	0.037	0.034	0.033	0.033	0.032	0.042	0.035	0.010
	Long	5%	0.065	0.068	0.071	0.027	0.025	0.059	0.016	0.057	0.004
		15%	0.038	0.037	0.039	0.017	0.023	0.033	0.009	0.035	0.002
2PL	Short	5%	0.070	0.069	0.050	0.049	0.048	0.058	0.056	0.067	0.032
		15%	0.041	0.038	0.028	0.029	0.033	0.033	0.032	0.035	0.018
	Long	5%	0.062	0.068	0.056	0.014	0.022	0.063	0.016	0.051	0.020
		15%	0.036	0.038	0.035	0.007	0.021	0.036	0.008	0.035	0.011

Note: EL = exam length. SA = sample aberrance.

Table 32: *Sensitivity - Mean Values for Combined Person-Fit Measures Under Low-Level T Conditions*

			Theta level: Low											
			Exam aberrance											
			10%				25%				50%			
Model	EL	SA	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT
Rasch	Short	5%	0.722	0.697	0.631	0.782	0.986	0.905	0.916	0.987	0.941	0.818	0.675	0.941
		15%	0.713	0.695	0.620	0.773	0.980	0.904	0.882	0.980	0.908	0.818	0.513	0.908
	Long	5%	0.962	0.960	0.942	0.976	1.000	0.985	0.993	1.000	0.999	0.989	0.836	0.999
		15%	0.959	0.960	0.938	0.975	0.999	0.985	0.987	0.999	0.997	0.989	0.688	0.997
2PL	Short	5%	0.791	0.739	0.495	0.812	0.994	0.949	0.916	0.994	0.981	0.941	0.739	0.981
		15%	0.790	0.740	0.489	0.810	0.991	0.949	0.888	0.991	0.971	0.941	0.606	0.971
	Long	5%	0.962	0.943	0.767	0.967	1.000	0.997	0.995	1.000	1.000	1.000	0.898	1.000
		15%	0.958	0.943	0.760	0.964	1.000	0.997	0.989	1.000	1.000	1.000	0.799	1.000

Note: EL = exam length. SA = sample aberrance.

Table 33: *Sensitivity – Standard Deviation of the Mean for Combined Person-Fit Measures Under Low-Level T Conditions*

			Theta level: Low											
			Exam aberrance											
			10%				25%				50%			
Model	EL	SA	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT
Rasch	Short	5%	0.063	0.068	0.073	0.062	0.016	0.041	0.039	0.016	0.034	0.055	0.065	0.034
		15%	0.037	0.039	0.039	0.036	0.012	0.025	0.025	0.012	0.023	0.032	0.037	0.023
	Long	5%	0.027	0.027	0.033	0.022	0.003	0.017	0.012	0.003	0.005	0.015	0.049	0.005
		15%	0.015	0.016	0.019	0.012	0.002	0.010	0.009	0.002	0.005	0.009	0.033	0.005
2PL	Short	5%	0.057	0.063	0.071	0.057	0.011	0.032	0.040	0.011	0.018	0.032	0.060	0.018
		15%	0.032	0.035	0.040	0.032	0.007	0.018	0.026	0.007	0.013	0.020	0.039	0.013
	Long	5%	0.025	0.031	0.059	0.024	0.001	0.008	0.010	0.001	0.001	0.002	0.042	0.001
		15%	0.015	0.018	0.035	0.015	0.001	0.005	0.008	0.001	0.001	0.001	0.031	0.001

Note: EL = exam length. SA = sample aberrance.

Table 34: *Sensitivity – Mean Values for Combined Person-Fit Measures Under Mid-Level T Conditions*

			Theta level: Mid											
			Exam aberrance											
			10%				25%				50%			
Model	EL	SA	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT
Rasch	Short	5%	0.722	0.554	0.535	0.740	0.973	0.784	0.898	0.973	0.841	0.610	0.591	0.841
		15%	0.697	0.553	0.494	0.716	0.942	0.786	0.782	0.942	0.761	0.608	0.386	0.761
	Long	5%	0.863	0.700	0.778	0.888	0.999	0.961	0.977	0.999	0.997	0.987	0.775	0.997
		15%	0.840	0.701	0.736	0.867	0.996	0.961	0.919	0.996	0.995	0.987	0.547	0.995
2PL	Short	5%	0.734	0.530	0.495	0.743	0.982	0.851	0.902	0.983	0.917	0.782	0.618	0.917
		15%	0.707	0.528	0.454	0.718	0.961	0.850	0.790	0.961	0.870	0.786	0.404	0.870
	Long	5%	0.886	0.698	0.681	0.892	1.000	0.991	0.983	1.000	0.998	0.987	0.819	0.998
		15%	0.869	0.704	0.633	0.876	0.999	0.991	0.940	0.999	0.995	0.988	0.598	0.995

Note: EL = exam length. SA = sample aberrance.

Table 35: *Sensitivity – Standard Deviation of the Mean for Combined Person-Fit Measures Under Mid-Level T Conditions*

			Theta level: Mid											
			Exam aberrance											
			10%				25%				50%			
Model	EL	SA	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT
Rasch	Short	5%	0.063	0.069	0.071	0.062	0.023	0.057	0.041	0.023	0.052	0.070	0.068	0.052
		15%	0.038	0.040	0.040	0.036	0.019	0.033	0.031	0.019	0.034	0.042	0.035	0.034
	Long	5%	0.046	0.064	0.056	0.041	0.004	0.027	0.021	0.004	0.008	0.016	0.057	0.008
		15%	0.028	0.036	0.033	0.026	0.005	0.017	0.021	0.005	0.006	0.009	0.035	0.006
2PL	Short	5%	0.062	0.071	0.070	0.062	0.018	0.049	0.040	0.018	0.039	0.056	0.066	0.039
		15%	0.038	0.041	0.039	0.037	0.015	0.029	0.032	0.015	0.026	0.032	0.037	0.026
	Long	5%	0.043	0.062	0.065	0.042	0.001	0.014	0.018	0.001	0.006	0.016	0.051	0.006
		15%	0.027	0.036	0.038	0.027	0.002	0.007	0.018	0.002	0.006	0.008	0.035	0.006

Note: EL = exam length. SA = sample aberrance.

Table 36: *Specificity – Mean Values for Individual Person-Fit Measures Under Low-Level T Conditions*

			Theta level: Low								
			Exam aberrance								
			10%			25%			50%		
Model	EL	SA	l_z	ERT	PRC	l_z	ERT	PRC	l_z	ERT	PRC
Rasch	Short	5%	0.949	0.897	0.967	0.949	0.904	0.967	0.949	0.909	0.967
		15%	0.946	0.914	0.970	0.946	0.930	0.970	0.946	0.940	0.970
	Long	5%	0.953	0.919	0.871	0.953	0.928	0.871	0.953	0.934	0.871
		15%	0.959	0.935	0.897	0.959	0.953	0.897	0.959	0.964	0.897
2PL	Short	5%	0.950	0.899	0.970	0.950	0.905	0.970	0.950	0.909	0.970
		15%	0.949	0.918	0.973	0.949	0.932	0.973	0.949	0.940	0.973
	Long	5%	0.953	0.908	0.926	0.953	0.916	0.926	0.953	0.921	0.926
		15%	0.960	0.931	0.943	0.960	0.947	0.943	0.960	0.955	0.943

Note: EL = exam length. SA = sample aberrance.

Table 37: *Specificity – Standard Deviation of the Mean for Individual Person-Fit Measures Under Low-Level T Conditions*

			Theta level: Low								
			Exam aberrance								
			10%			25%			50%		
Model	EL	SA	l_z	ERT	PRC	l_z	ERT	PRC	l_z	ERT	PRC
Rasch	Short	5%	0.007	0.009	0.006	0.007	0.009	0.006	0.007	0.008	0.006
		15%	0.008	0.009	0.006	0.008	0.008	0.006	0.008	0.008	0.006
	Long	5%	0.007	0.008	0.010	0.007	0.008	0.010	0.007	0.007	0.010
		15%	0.007	0.008	0.010	0.007	0.007	0.010	0.007	0.006	0.010
2PL	Short	5%	0.007	0.008	0.006	0.007	0.008	0.006	0.007	0.008	0.006
		15%	0.008	0.008	0.005	0.008	0.008	0.005	0.008	0.008	0.005
	Long	5%	0.006	0.008	0.007	0.006	0.008	0.007	0.006	0.007	0.007
		15%	0.006	0.008	0.007	0.006	0.007	0.007	0.006	0.006	0.007

Note: EL = exam length. SA = sample aberrance.

Table 38: *Specificity – Mean Values for Individual Person-Fit Measures Under Mid-Level T Conditions*

			Theta level: Mid								
			Exam aberrance								
			10%			25%			50%		
Model	EL	SA	l_z	ERT	PRC	l_z	ERT	PRC	l_z	ERT	PRC
Rasch	Short	5%	0.952	0.890	0.966	0.952	0.895	0.966	0.951	0.898	0.966
		15%	0.954	0.891	0.967	0.954	0.900	0.967	0.953	0.905	0.967
	Long	5%	0.950	0.914	0.856	0.950	0.920	0.856	0.950	0.923	0.856
		15%	0.949	0.916	0.849	0.949	0.927	0.849	0.949	0.930	0.850
2PL	Short	5%	0.951	0.891	0.968	0.951	0.896	0.968	0.951	0.898	0.968
		15%	0.953	0.892	0.968	0.953	0.901	0.968	0.952	0.903	0.968
	Long	5%	0.950	0.898	0.916	0.950	0.905	0.916	0.949	0.907	0.916
		15%	0.947	0.900	0.910	0.947	0.912	0.910	0.947	0.913	0.910

Note: EL = exam length. SA = sample aberrance.

Table 39: *Specificity – Standard Deviation of the Mean for Individual Person-Fit Measures Under Mid-Level T Conditions*

			Theta level: Mid								
			Exam aberrance								
			10%			25%			50%		
Model	EL	SA	l_z	ERT	PRC	l_z	ERT	PRC	l_z	ERT	PRC
Rasch	Short	5%	0.007	0.009	0.006	0.007	0.009	0.006	0.007	0.009	0.006
		15%	0.007	0.010	0.006	0.007	0.010	0.006	0.007	0.009	0.006
	Long	5%	0.007	0.008	0.010	0.007	0.008	0.010	0.007	0.008	0.010
		15%	0.007	0.008	0.011	0.007	0.008	0.011	0.007	0.007	0.011
2PL	Short	5%	0.007	0.008	0.006	0.007	0.009	0.006	0.007	0.008	0.006
		15%	0.007	0.009	0.006	0.007	0.009	0.006	0.007	0.009	0.006
	Long	5%	0.007	0.008	0.008	0.007	0.008	0.008	0.007	0.008	0.008
		15%	0.007	0.008	0.009	0.007	0.007	0.009	0.007	0.008	0.009

Note: EL = exam length. SA = sample aberrance.

Table 40: *Specificity – Mean Values for Combined Person-Fit Measures Under Low-Level T Conditions*

			Theta level: Low											
			Exam aberrance											
			10%				25%				50%			
Model	EL	SA	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT
Rasch	Short	5%	0.881	0.928	0.898	0.862	0.888	0.928	0.905	0.868	0.892	0.928	0.909	0.873
		15%	0.885	0.928	0.907	0.868	0.900	0.928	0.922	0.882	0.910	0.928	0.933	0.892
	Long	5%	0.894	0.854	0.818	0.803	0.903	0.854	0.826	0.811	0.909	0.854	0.832	0.817
		15%	0.909	0.881	0.851	0.836	0.926	0.881	0.868	0.853	0.937	0.881	0.878	0.863
2PL	Short	5%	0.887	0.931	0.906	0.870	0.893	0.931	0.912	0.875	0.897	0.931	0.915	0.879
		15%	0.894	0.933	0.916	0.878	0.907	0.933	0.931	0.892	0.915	0.933	0.938	0.899
	Long	5%	0.895	0.900	0.871	0.846	0.904	0.900	0.879	0.854	0.909	0.900	0.884	0.859
		15%	0.913	0.919	0.899	0.876	0.929	0.919	0.915	0.891	0.937	0.919	0.922	0.898

Note: EL = exam length. SA = sample aberrance.

Table 41: *Specificity – Standard Deviation of the Mean for Combined Person-Fit Measures Under Low-Level T Conditions*

			Theta level: Low											
			Exam aberrance											
			10%				25%				50%			
Model	EL	SA	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT
Rasch	Short	5%	0.010	0.008	0.009	0.010	0.010	0.008	0.008	0.010	0.009	0.008	0.008	0.010
		15%	0.011	0.008	0.009	0.011	0.010	0.008	0.009	0.011	0.009	0.008	0.008	0.010
	Long	5%	0.009	0.011	0.011	0.012	0.009	0.011	0.011	0.012	0.009	0.011	0.011	0.012
		15%	0.010	0.011	0.012	0.012	0.009	0.011	0.011	0.012	0.008	0.011	0.011	0.011
2PL	Short	5%	0.010	0.008	0.009	0.011	0.010	0.008	0.009	0.011	0.010	0.008	0.008	0.010
		15%	0.010	0.009	0.009	0.011	0.010	0.009	0.008	0.011	0.010	0.009	0.008	0.010
	Long	5%	0.009	0.009	0.010	0.011	0.009	0.009	0.010	0.010	0.009	0.009	0.009	0.010
		15%	0.009	0.009	0.010	0.011	0.008	0.009	0.009	0.010	0.008	0.009	0.008	0.009

Note: EL = exam length. SA = sample aberrance.

Table 42: *Specificity – Mean Values for Combined Person-Fit Measures Under Mid-Level T Conditions*

			Theta level: Mid											
			Exam aberrance											
			10%				25%				50%			
Model	EL	SA	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT
Rasch	Short	5%	0.883	0.929	0.897	0.862	0.887	0.929	0.901	0.867	0.889	0.929	0.904	0.869
		15%	0.890	0.932	0.902	0.869	0.899	0.932	0.911	0.878	0.901	0.931	0.916	0.881
	Long	5%	0.890	0.840	0.803	0.789	0.896	0.840	0.810	0.795	0.899	0.840	0.812	0.797
		15%	0.894	0.833	0.802	0.787	0.905	0.833	0.812	0.797	0.908	0.834	0.816	0.801
2PL	Short	5%	0.888	0.931	0.903	0.868	0.892	0.931	0.908	0.873	0.894	0.930	0.910	0.875
		15%	0.895	0.932	0.909	0.875	0.903	0.932	0.918	0.883	0.904	0.931	0.920	0.885
	Long	5%	0.889	0.889	0.859	0.834	0.896	0.889	0.865	0.840	0.898	0.889	0.867	0.842
		15%	0.893	0.882	0.859	0.833	0.904	0.882	0.870	0.844	0.906	0.882	0.872	0.846

Note: EL = exam length. SA = sample aberrance.

Table 43: *Specificity – Standard Deviation of the Mean for Combined Person-Fit Measures Under Mid-Level T Conditions*

			Theta level: Mid											
			Exam aberrance											
			10%				25%				50%			
Model	EL	SA	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT	l_z + ERT	l_z + PRC	PRC+ ERT	l_z + PRC+ ERT
Rasch	Short	5%	0.010	0.008	0.009	0.010	0.010	0.008	0.008	0.010	0.010	0.008	0.009	0.010
		15%	0.010	0.009	0.009	0.011	0.010	0.009	0.009	0.011	0.010	0.009	0.009	0.011
	Long	5%	0.009	0.011	0.012	0.012	0.009	0.011	0.012	0.012	0.009	0.011	0.012	0.012
		15%	0.010	0.012	0.013	0.013	0.009	0.012	0.012	0.013	0.009	0.012	0.012	0.013
2PL	Short	5%	0.010	0.008	0.009	0.011	0.010	0.008	0.009	0.011	0.010	0.008	0.009	0.010
		15%	0.011	0.009	0.009	0.011	0.010	0.009	0.009	0.011	0.010	0.008	0.009	0.011
	Long	5%	0.010	0.009	0.010	0.011	0.009	0.009	0.010	0.011	0.009	0.009	0.010	0.011
		15%	0.010	0.010	0.011	0.011	0.009	0.010	0.010	0.011	0.009	0.010	0.010	0.011

Note: EL = exam length. SA = sample aberrance.

CHAPTER 5: DISCUSSION AND CONCLUSIONS

With today's computer-based testing systems, researchers and test administrators have access to detailed item response and response time data, providing the opportunity to learn more about how examinees interact with a test. These kinds of data can also be used to ensure that the integrity of exam scores, and the testing program as a whole, is maintained. As practitioners, it is important to keep in mind that the consequences of asserting that test scores may have been the result of cheating can be severe for both the examinee and the test administrator. Using multiple indicators to gather evidence when making a determination about aberrant response behavior is a promising alternative to relying on a single measure.

The purpose of this study was to evaluate the use of multiple person-fit indices in determining whether aberrant (cheating) behavior was exhibited on an exam. This research explored whether using indices that evaluate various aspects of a response pattern (likelihood of the response pattern, shape of the curve, response time) increased the accuracy in classifying a response pattern as aberrant. In addition, the impact of various factors on the effectiveness of person-fit indices was explored.

As discussed in Chapter 2, there are numerous person-fit indices that have been developed and well researched. Determining which are best for a given exam program depends on many factors, such as measurement model employed, exam length, and item type(s), to name a few. The decision of which person-fit indices are most appropriate for use is largely left to the test administrator as there is no general consensus in the literature. As such, researchers and test administrators should conduct their own studies, drawing from the work of others, to gather empirical data on how best to implement various person-fit indices in an exam program. To aid in this process, Rupp (2013) provided practical recommendations to consider when designing a

person-fit simulation study to increase the likelihood that results would be of use in an operational setting. Rupp's recommendations were incorporated into the present study.

A real-parameter simulation design was employed for this study. This entailed using Rasch and 2PL item and person parameters estimated using data from an operational exam program to simulate baseline data. These data were then manipulated to simulate cheating behavior in order to evaluate three person-fit indices, alone and in combination, in the identification of aberrant response patterns. As recommended by Seo and Weiss (2013), the baseline data were also used to determine the appropriate l_z critical values for flagging aberrance. This method proved useful as the critical values for the 40-item (short) exam were considerably different from those for the 100-item (long) exam. The average critical values across simulation replications for the short exam were -1.624 for the Rasch model and -1.684 for the 2PL model, both of which are close to the theoretical cut of -1.645 at $\alpha = .05$. The average critical values for the long exam were -2.547 for the Rasch model and -2.376 for the 2PL model.

Impact of Study Conditions on Detecting Aberrance

The results of the ANOVA showed that l_z performed consistently better than ERT and PRC, producing small to large effect sizes and lower Type I error rates across various study conditions. The ERT showed promise as a means for evaluating response time information as an indication of aberrant responding. However, there was some uncertainty as to whether these findings were an artifact of reducing response time by a constant (50%) for all items selected to mimic cheating. This should be investigated further using alternative simulation designs. As described by Wang and Xu (2015) and Wang, Xu, and Shang (2016), a response time distribution modeled for cheating behavior can be utilized rather than systematically reducing response time by a constant.

While PRC did not perform as well when the PRC χ^2 statistic alone was used to classify examinees, PRC can add value to the process for evaluating aberrant response patterns through the graphical representation of the observed and expected person response curves. This was demonstrated in Chapter 3, Figures 1 – 4. Additional examples of the use of the PRC can be seen in Appendix A.

Overall, exam length and percentage of exam aberrance had the largest effect on the person-fit measures. All three measures performed best on the long form and when exam aberrance was set to 25%. In particular, l_z produced Type I error rates of 0.017 and 0.009 for the long form at 25% and 50% exam aberrance, respectively. Analyses also revealed that l_z performed better for low theta levels on the long form, with Type I error rates of 0.009 and 0.006 for the low-level theta, long form, 25% exam aberrance condition. The ERT Type I error rates, while above an acceptable range, were also lowest for the 25% exam aberrance condition. It is also worth noting that, in general, the person-fit measures performed better when exam aberrance was set to 50% than 10%, but clearly not as well as when exam aberrance was at 25%. These findings may indicate that aberrance on only a few items will not flag an examinee as an aberrant responder as the aberrance could be caused by things like carelessness or lucky guessing. Alternatively, with higher levels of aberrant responding on an exam, it may be harder to discern whether a low ability examinee had access to content prior to the exam or if a high ability examinee is just making a lot of mistakes due to carelessness or fatigue.

Interaction effect sizes for exam length and exam aberrance, exam aberrance and theta level, and the combination of all three factors produced small effect sizes with the individual measure l_z. A small effect size was also present with ERT under the exam aberrance by theta condition. Not surprising, l_z + ERT and l_z + ERT + PRC produced the largest effect size (0.25) across all interaction effects for the exam length by exam

aberrance condition. The Type I error rates were also below nominal levels for l_z + ERT and l_z + ERT + PRC at the 25% exam aberrance condition (0.024 and 0.001 for short and long forms, respectively) and also for the l_z + ERT, l_z + PRC, and l_z + ERT + PRC for the long form, 50% exam aberrance condition (0.003, 0.009, and 0.003, respectively). Adding the theta-level condition (exam length by exam aberrance by theta level) further revealed that l_z + ERT and l_z + ERT + PRC performed well for examinees in the short form, 50% exam aberrance, low-level theta condition, with Type I error rates at 0.050. Finally, l_z + ERT and l_z + ERT + PRC produced the lowest Type I error rates (< 0.001) across conditions within the long form, 25% exam aberrance, low-level theta condition.

In summary, the results of the ANOVA showed that l_z consistently performed the best across study conditions and that l_z + ERT added value particularly when exam aberrance was 25% or more. These measures also performed better with lower-level theta values. While the combination of all three person-fit measures produced positive results under the study conditions described above, PRC on its own did not produce notable results.

Classification Accuracy

Coefficient kappa values were generally low across person-fit measures. The highest values of kappa occurred between l_z and ERT under the long form, 25% exam aberrance conditions. Average kappa values ranged from 68% to 75% agreement. Additionally, l_z and PRC exhibited percentages of agreement between 65% and 72% for the long exam, 15% sample aberrance, low-level theta conditions. This was not unexpected since, based on the ANOVA results, l_z has proven to be a more consistent and reliable measure of person fit.

Reviewing the numeric kappa values alone did not prove very useful in drawing conclusions about the performance of the person-fit indices. Graphing the results

provided confirmation of the ANOVA results, namely that l_z and ERT are generally the most consistent with regard to identifying aberrant response patterns.

Gathering multiple sources of empirical evidence is critical when making judgements about aberrance, especially cheating behavior. Not only is consistency in classification important, but the incremental value in utilizing multiple measures and how this impacts the decision-making process is of equal, if not higher, importance. In evaluating classification accuracy and diagnostic efficiency, sensitivity and specificity are two measures used to assess how often correct decisions are made.

A perfect person-fit measure would correctly discriminate between cheaters and non-cheaters 100% of the time. Unfortunately, this is not a realistic scenario, especially given the unpredictable nature of cheating on tests. As discussed in Chapters 1 and 2, the cost of labeling an examinee as cheating on an exam must be seriously considered before making a final decision. Sensitivity and specificity provide measures of decision accuracy. That is, sensitivity provides an estimate of the probability that a person-fit measure will correctly identify an examinee who is actually cheating, whereas specificity provides an estimate of the probability that the person-fit measure will correctly identify an examinee who is not cheating.

The results of the sensitivity analysis performed for the present study showed that l_z on its own provided reasonable levels of sensitivity across study conditions. Sensitivity values were consistently above 0.961 across long form conditions, reaching 1.000 under the 2PL model, low-level theta, 50% exam aberrance condition. Sensitivity values were generally lower for the mid-level theta condition, especially under the 10% exam aberrance condition in which values ranged from 0.514 to 0.692. Sensitivity values using multiple measures were promising, especially for low-level theta conditions. Sensitivity values for l_z + ERT for the 25% exam aberrance condition ranged from 0.942 to 1.000 across all other conditions. It should be noted that the values for l_z + PRC and

ERT + PRC were also very high across the 25% exam aberrance conditions, especially for the low-level theta conditions. An interesting finding was that l_z + ERT + PRC values matched those for l_z + ERT under the 25% and 50% exam aberrance conditions. Because the sensitivity values for PRC alone were considerably lower than the values for l_z and ERT (individually), it was concluded that PRC may not be adding to the to the sensitivity of the combined measures. Rather, the values for l_z + ERT + PRC could merely be a reflection of the value of using l_z and ERT in combination to flag aberrant responding.

The results of the specificity analysis were very encouraging. Values for the individual person-fit indices were high across all study conditions. For the low-level theta conditions, specificity values were generally above 0.900, with the lowest value produced by PRC under the Rasch model, long form, 5% sample aberrance conditions. Specificity values across the mid-level theta conditions fluctuated slightly, however the lowest value was still relatively high at 0.849 for PRC under the Rasch model, long form 15% conditions.

Specificity values for the combined measures were slightly lower overall than the values for the individual measures. This indicates that the person-fit measures are better at classifying true non-cheaters as non-cheaters on their own, rather than in combination with another measure. This is supported by the fact that the lowest specificity value of 0.803 was produced by l_z + ERT + PRC within the low-level theta condition. The highest value within the low-level theta condition was 0.938, produced by ERT + PRC. Similarly, the lowest specificity value within the mid-level theta condition was 0.787, produced by l_z + ERT + PRC, and the highest value was 0.932, produced by l_z + PRC.

Conclusions

The factors that had the most impact on the person-fit measures in terms of identifying aberrant response patterns were exam length and percentage of aberrance introduced at the exam level. The factor that had the least impact was the percentage of aberrance introduced in a given sample of examinees. In real life settings, exam aberrance and sample aberrance are typically unknown and cannot be controlled by the test administrator (outside of taking precautionary measures to minimize item exposure, ensure security of exam content, and implement policies to deter cheating). Exam length, however, is a factor that can be controlled by the test administrator and should be a consideration when cheating is a concern, provided a long exam is supported by the test blueprint and can be tolerated by the population being assessed.

Although exam aberrance cannot be controlled by the test administrator, it is important to keep in mind that the measures used in this study were most effective under the 25% exam aberrance condition. Exam aberrance at the extremes (less than 10% or greater than 50%) may prove more difficult to detect.

With regard to theta level, aberrant responders in the low-level theta condition were identified more often than those in the mid-level theta condition. This may occur because an exam typically contains a higher percentage of items with a difficulty level above the ability level of lower ability examinees. This means that examinees with lower ability levels have more opportunity to exhibit aberrant response patterns, especially on longer exams, simply because there are more items that they are expected to answer wrong. Examinees achieving high scores, and therefore high ability estimates due to cheating behavior are harder to identify because there are fewer response inconsistencies to evaluate. That is, high ability examinees may have answered a few easy items incorrectly, but overall answered the majority of the easy and hard items correctly. The easy items that were missed may appear as carelessness rather than an

attempt to cheat since the overall proportion of unexpected responses in low. In addition, faster item response times are often not unexpected with high ability examinees.

While the PRC χ^2 did not prove to be an effective means for identifying aberrant response patterns in this study, use of the PRC as a graphical tool should not be ruled out. As shown in Figures 2 and 4, use of the PRC as a graphical means for further investigation and understanding of aberrance can be of added value. These Figures allow for visual inspection of the observed PRC compared to the expected PRC. As can be seen in Figure 2, the responses of a low ability examinee on the low difficulty items appear to be somewhat aberrant. Conversely, Figure 4 shows the responses of a high ability examinee as appearing quite aberrant throughout much of the exam.

The results of this study did not reveal a notable impact of IRT model on the performance of the person-fit indices. This may be due, in part, to the fact that the person-fit measures are more heavily influenced by item difficulty levels. Further research is needed to explore whether these or other person-fit measures are sensitive to varying levels of item discrimination. Identifying such a measure could add a new dimension to detecting aberrant responding.

The results of the diagnostic efficiency analysis were encouraging. Tables 44 and 45 below were added to provide concise recommendations regarding the person-fit measures that were most effective in providing high levels of sensitivity and specificity. Because it is prudent to be conservative when classifying examinees as exhibiting cheating behavior, the recommendations below are based on sensitivity and specificity values > 0.900 (unless otherwise noted). It was found that using multiple person-fit measures provided the best classification accuracy in terms of sensitivity and therefore, Table 44 only contains recommendations for combined measures. It should be noted that there are several conditions in which no set of combined measures was

recommended. In these instances, there were no individual or combined measures that provided sensitivity values greater than 0.900. However, values were generally between 0.695 and 0.886 for l_z + ERT under these conditions and therefore, could be used to estimate sensitivity. In addition, high sensitivity values were noted for l_z and ERT (on an individual basis) under conditions identified by the ANOVA to have the greatest impact on the ability of person-fit indices to detect cheating.

Table 44: *Sensitivity – Recommended Combined Measures, by Study Factors, with Values > 0.900*

		Theta Level: Low			Theta Level: Mid		
		Exam aberrance			Exam aberrance		
Model	EL	10%	25%	50%	10%	25%	50%
Rasch	Short	N/A	All[a]	l_z+ERT	N/A	l_z+ERT	N/A
	Long	All	All	l_z+ERT, l_z+PRC[b]	N/A	l_z+ERT, l_z+PRC, ERT+PRC[b]	l_z+ERT, l_z+PRC[b]
2PL	Short	N/A	All[a]	l_z+ERT, l_z+PRC[b]	N/A	l_z+ERT, l_z+ERT +PRC	l_z+ERT[bc]
	Long	l_z+ERT, l_z+PRC, l_z+ERT +PRC	All	l_z+ERT, l_z+PRC[b]	N/A	l_z+ERT, l_z+PRC, ERT+PRC[b]	l_z+ERT, l_z+PRC[b]

Note: EL = exam length. Cells containing 'All' signify all combinations of individual measures (i.e., l_z+ERT, l_z+PRC, ERT+PRC, l_z+ERT+PRC) had sensitivity values greater than 0.900, except where noted.

[a]ERT+PRC >= 0.882 for 15% SA condition

[b]l_z+ERT+PRC is not included because sensitivity values are the same as l_z+ERT

[c]l_z+ERT = 0.870

Unlike the findings for sensitivity, specificity values were found to be highest when using individual person-fit measures. As such, Table 45 only contains recommendations for individual measures with specificity values > 0.900 (unless otherwise noted).

Table 45: *Specificity – Recommended Individual Measures by Study Factors, with Values > 0.900*

		Theta Level: Low			Theta Level: Mid		
		Exam aberrance			Exam aberrance		
Model	EL	10%	25%	50%	10%	25%	50%
Rasch	Short	All[a]	All	All	l_z, PRC	All[a]	All[a]
	Long	l_z, ERT	l_z, ERT	l_z, ERT	l_z, ERT	l_z, ERT	l_z, ERT
2PL	Short	All[a]	All	All	l_z, PRC	All[a]	All[a]
	Long	All	All	All	All[a]	All	All

Note: EL = exam length. Cells containing 'All' signify all individual measures (i.e., l_z, ERT, and PRC) had specificity values greater than 0.900, except where noted.
[a]ERT >= 0.895 for 5% SA condition

When cheating is a concern on an exam, administering longer exams (e.g., 100 items) can provide test administrators with more reliable information on which to make decisions regarding the classification of examinees. Our goal as test administrators is to provide examinees with a fair opportunity to demonstrate competency and ability. Decisions regarding the classification of examinees as exhibiting cheating or aberrant behavior on a test and what action will be taken based on this information should be determined only when empirical data can be taken into account to ensure all examinees are treated fairly and equally.

The results of studies like the present one can be used when designing a new exam and/or updating an existing exam. This type of information can also be used to take steps to increase the overall security of an exam program. This type of analysis can shed light on how well current processes are working and also provide information on where to focus future development efforts or policy changes, such as the amount of time an examinee must wait to retest. There are many actions that can be taken based on the results of these types of inquiries. The consequences of classifying someone as

exhibiting cheating behavior must be carefully weighed with factors like the risk to public protection and maintaining the integrity of an exam program.

Future Considerations

Although there has been considerable research done in the area of detecting aberrant responding, additional work is needed to develop reliable and practical methods. There are several types of person-fit indices that were not addressed in this study. For example, the optimal person-fit statistics have been found to be the most powerful tests available of the null hypothesis. Future research might focus on these statistics and how practitioners could efficiently implement one or more within an operational exam program.

The ERT measure shows promise, but there is limited research available on the implementation and use of the ERT to detect aberrant responding. In general, the focus of much of the research on item response time has been in identifying carelessness and random responding. Test delivery technology is advancing and new features, such as tracking the number of times an examinee goes back to review one or more items and the amount of time spent reviewing those items, are being developed. With the availability of such detailed response time information in computer-based testing, there is much that can be learned regarding cheating behavior. ERT may be the first of many response time person-fit indices that can be used to hone in on the detection of aberrant response patterns.

The ERT measure should also be evaluated using alternative simulation designs in which the item response times for examinees designated as cheaters are drawn from a distribution modeled for aberrant response behavior. This would provide information regarding whether the effectiveness of ERT found in this study was the result of utilizing a deterministic simulation design (i.e., reducing item response time by 50% for a specified percentage of items on an exam).

Other areas of future research might include a narrower focus in regard to study conditions. A total of 48 conditions were included in this research. It would be beneficial for studies to focus on a subset of factors and possibly gain a more in-depth understanding of how best to detect aberrance under certain conditions. Additionally, more work is needed using the 2PL model to understand whether item discrimination can offer more information, allowing for better detection of aberrant responding.

As noted throughout the study, exam length had an impact on the effectiveness of the person-fit measures and that the measures performed better for longer exams. While implementing longer exams is often feasible and typically utilized with fixed-length exams, this may not be an option for a computer-adaptive exam where the goal is to administer just enough items to obtain a reliable estimate of ability. Using multiple measures with a computer-adaptive exam may be an option, but requires more research.

Finally, the use of sensitivity and specificity in the current research helped to highlight the effectiveness of individual and combined measures from different viewpoints in identifying aberrant response patterns. Researchers might also consider using other measures of diagnostic efficiency, such as positive predictive power (PPP) and negative predictive power (NPP), when evaluating the discrimination ability of person-fit measures. In addition, the corrected form of kappa should be evaluated to determine whether it provides a better measure of classification accuracy than the uncorrected form of kappa.

The challenge of research in the area of aberrant responding (and particularly, cheating behavior) is that there isn't a measure of truth. That is, we never really know how many or which examinees have cheated on an exam and so simulation studies are employed in order to have a known target or "measure of truth". Using real-parameter simulations is one way to incorporate operational or practical data into a study. Finding

new ways to ensure simulation studies reflect real life will only enhance the ability of test administrators to maintain the security of their exams.

REFERENCES

Belov, D. I. (2013). Detection of test collusion via Kullback–Leibler divergence. *Journal of Educational Measurement, 50*(2), 141-163.

Brown, R. S. & Villarreal, J. C. (2007). Correcting for person misfit in aggregated score reporting. *International Journal of Testing, 7*, 1-25.

Cai, L. (2013). flexMIRT® version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. *Chapel Hill, NC: Vector Psychometric Group*.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Routledge.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37-46.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155.

Davenport Jr, E. C., & El-Sanhurry, N. A. (1991). Phi/phimax: review and synthesis. *Educational and Psychological Measurement, 51*(4), 821-828.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford.

de la Torre, J. & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*(2), 157-177.

DiSario, R., Olinsky, A., Quinn, J., & Schumacher, P. (2014). Applying Monte Carlo simulation to determine the likelihood of cheating on a multiple-choice professional exam. *Case Studies In Business, Industry And Government Statistics, 3(1), 30-36.*

Doyle, A. E., Biederman, J., Seidman, L. J., Weber, W., & Faraone, S. V. (2000). Diagnostic efficiency of neuropsychological test scores for discriminating boys

with and without attention deficit–hyperactivity disorder. *Journal of consulting and clinical psychology, 68*(3), 477.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67-86.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9*(1), 47-64.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Ferguson, C.J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532-538.

Ferrando, P. J., & Lorenzo, U. (2000). WPerfit: A program for computing parametric person-fit statistics and plotting person response curves. *Educational and Psychological Measurement*, *60*(3), 479-487.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology: General*, *141*(1), 2.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff Publishing.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement, 20*(2), 191-206.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied psychological measurement, 20*(2), 101-125.

Hauser, C., Kingsbury, G. G., & Houser, R. L. (2011, April). Individual Score Validity: Using the Wariness Index to Identify Test Performance to Treat with Caution. In *Presented at the Annual Meeting of the National Council on Measurement in Education*.

International Test Commission. (2014). International guidelines on the security of tests, examinations, and other assessments. Retrieved from www.intestcom.org

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277-298.

Klauer, K. C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56*(2), 213-228.

Klauer, K. C. (1995). The assessment of person fit. In *Rasch models* (pp. 97-110). Springer New York.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices (Third Edition)*. New York, NY: Springer.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35(1)*, 42-56.

Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53*(2), 161-176.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational and Behavioral Statistics, 4(4)*, 269-290.

Linacre, J. M. (2016). WINSTEPS Rasch measurement computer program. *Chicago: WINSTEPS.com*.

McLeod, L.D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23(2),* 147-160.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8*, 72-87.

Meijer, R. R. & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

Meijer, R. R., & Sotaridona, L. S. (2006). *Detection of advance item knowledge using response times in computer adaptive testing* (Vol. 3, No. 3). Law School Admission Council.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*(1), 75-106.

Molenaar, I. W., & Hoijtink, H. (1996). Person-fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education, 9*(1), 27-45.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*(2), 121-129.

Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*(2), 115-127.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14*, 127-137.

Reise, S. P. & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15*(3), 217-226.

Reise, S. P & Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. *Psychological Methods, 4*, 3-21.

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling, 55*(1), 3-38.

Schaeffer, G. A., Reese, C. M., Steffen, M., McKinley, R. L., & Mills, C. N. (1993). Field test of a computer-based GRE general test. ETS Research Report Series, 1993(1), i-47.

Schnipke, D. L., & Scrams, D. J. (1999). Exploring Issues of Test Taker Behavior: Insights Gained from Response-Time Analyses. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.

Seo, D. G., & Weiss, D. J. (2013). l_z person-fit index to identify misfit students with achievement test data. *Educational and Psychological Measurement, 73*(6), 994-1016.

Sijtsma, K. & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66*, 191-207.

Šimundić, A. M. (2008). Measures of diagnostic accuracy: basic definitions. *Medical and biological sciences, 22*(4), 61-65.

St-Onge, C., Valois, P., Abdous, B. & Germain, S. (2011). Accuracy of person-fit statistics: A Monte Carlo study of the influence of aberrance rates. *Applied Psychological Measurement, 35*, 419-432.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49*(1), 95-110.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7*(1), 81-96.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New Horizons in testing: Latent trait theory and computerized adaptive testing.* New York: Academic Press.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46*(3), 247-272.

van der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika, 68*(2), 251-265.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 327-345.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456-477.

Wang, C., Xu, G., & Shang, Z. (2016). A Two-Stage Approach to Differentiating Normal and Aberrant Behavior in Computer Based Testing. *Psychometrika*. doi:10.1007/s11336-016-9525-x

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19-38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183.

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010). An Investigation of the Relationship between Time of Testing and Test-Taking Effort. *Northwest Evaluation Association*.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97-116.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: Mesa Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

Zamost, S., Griffin, D., and Ansari, A. (2012, January 13). Prescription for cheating. *Cable News Network*. Retrieved from http://www.cnn.com/2012/01/13/health/prescription-for-cheating/

APPENDIX A: Person Response Curves (PRC) for Baseline and Manipulated Response Data by Study Condition

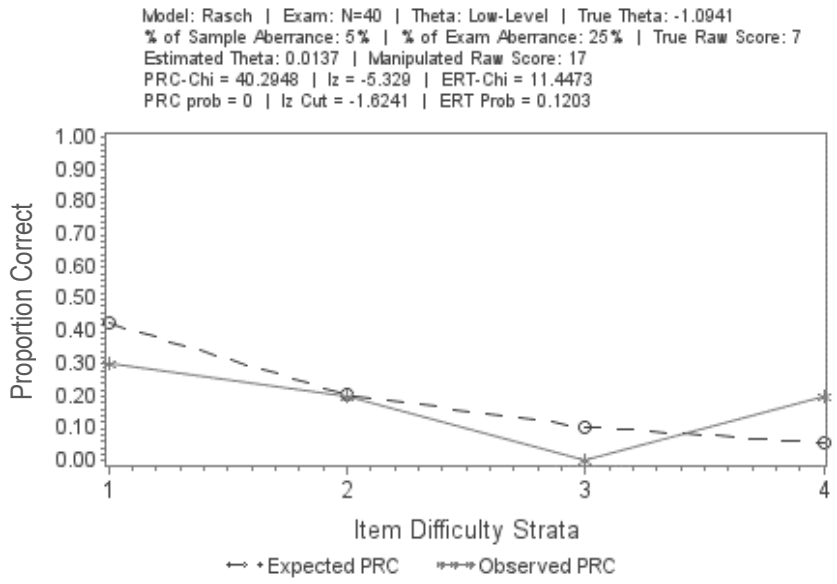

Figure A1. Baseline PRC for condition Rasch model x short form x low ability.

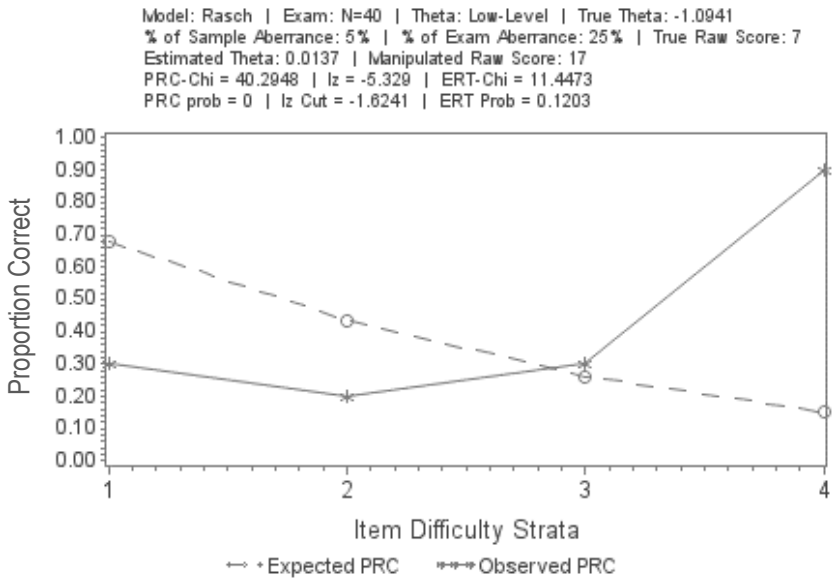

Figure A2. PRC for cheating condition: Rasch model x short form x low ability x 5% sample aberrance x 25% exam aberrance.

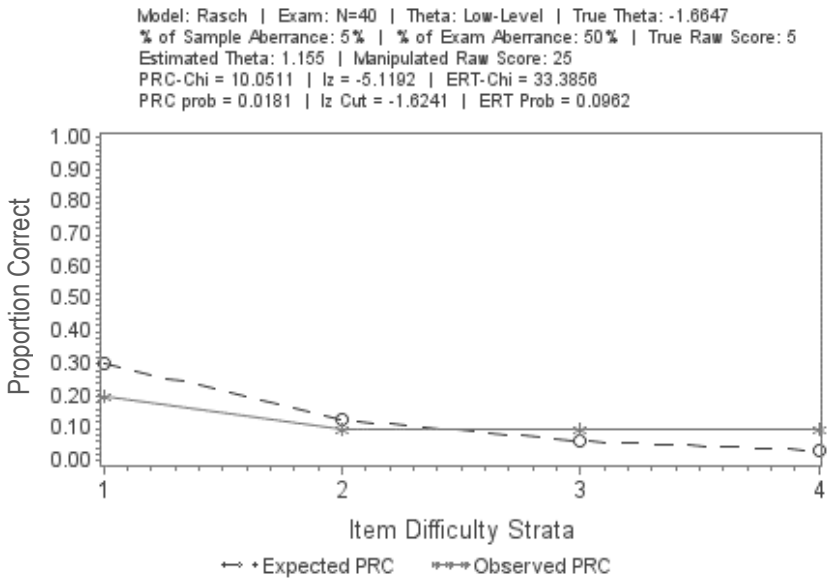

Figure A3. Baseline PRC for condition Rasch model x short form x low ability.

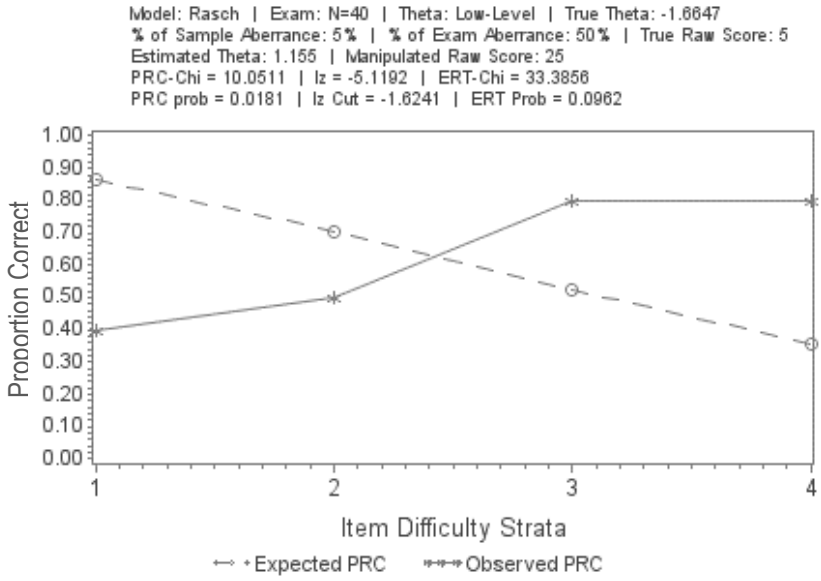

Figure A4. PRC for cheating condition: Rasch model x short form x low ability x 5% sample aberrance x 50% exam aberrance.

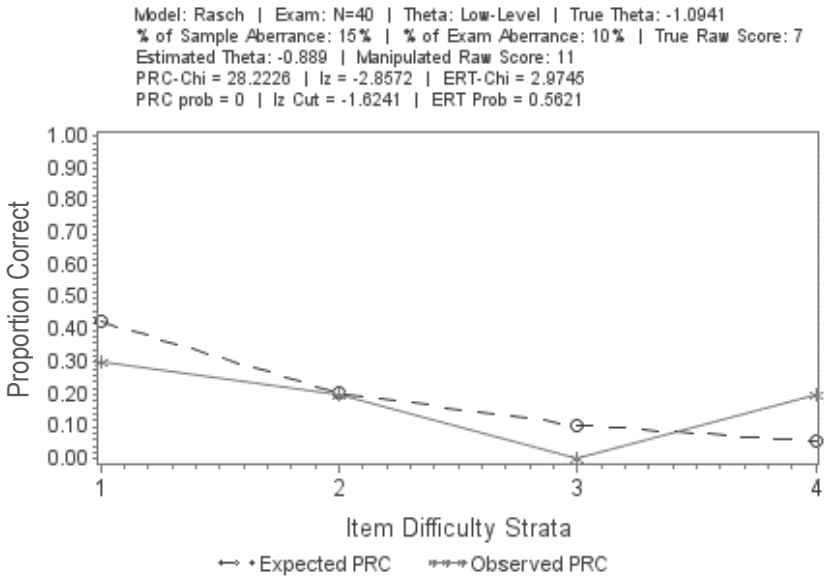

Figure A5. Baseline PRC for condition: Rasch model x short form x low ability.

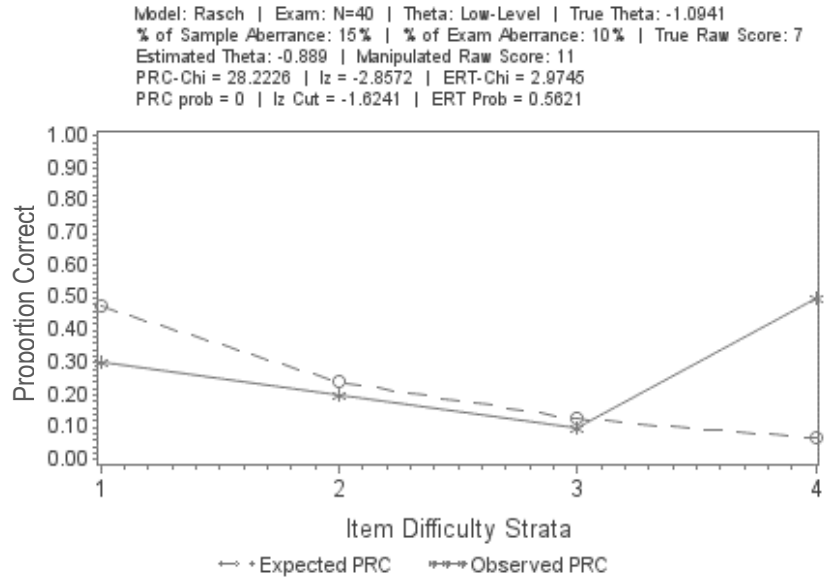

Figure A6. PRC for cheating condition: Rasch model x short form x low ability x 15% sample aberrance x 10% exam aberrance.

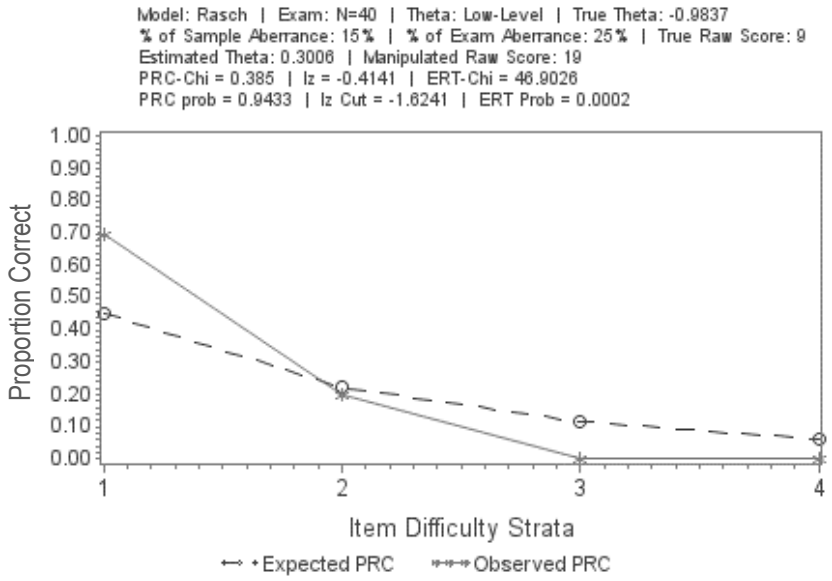

Figure A7. Baseline PRC for condition: Rasch model x short form x low ability.

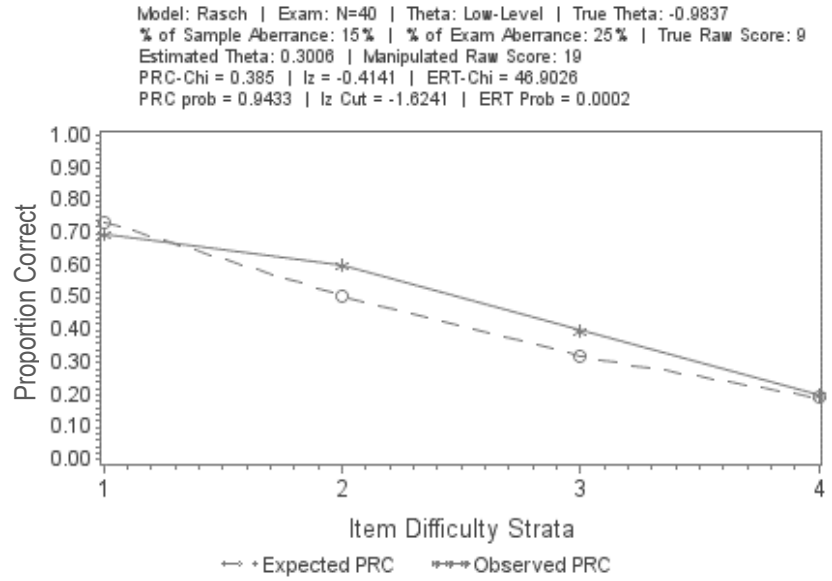

Figure A8. PRC for cheating condition: Rasch model x short form x low ability x 15% sample aberrance x 25% exam aberrance.

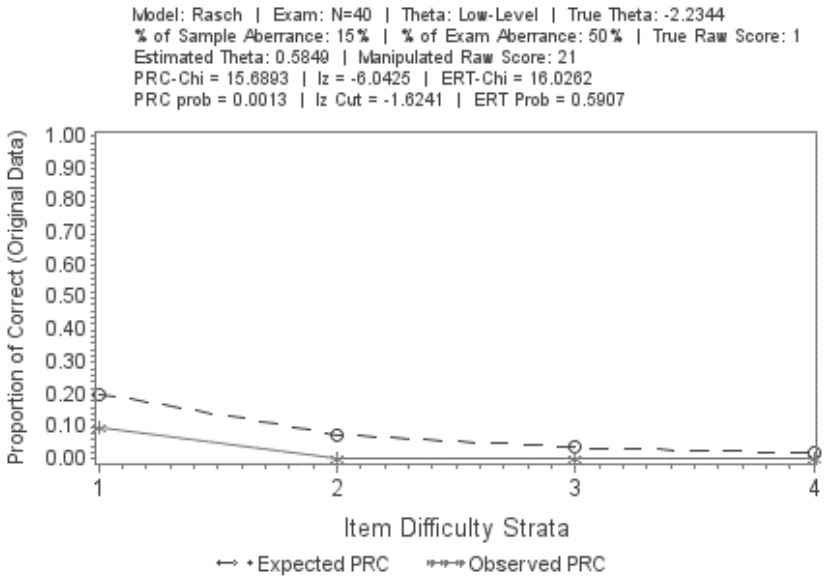

Figure A9. Baseline PRC for condition: Rasch model x short form x low ability.

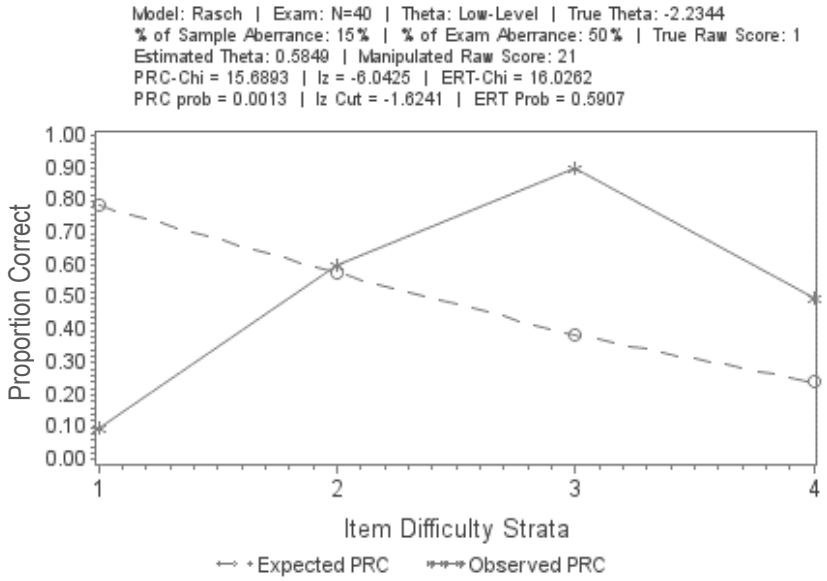

Figure A10. PRC for cheating condition: Rasch model x short form x low ability x 15% sample aberrance x 50% exam aberrance.

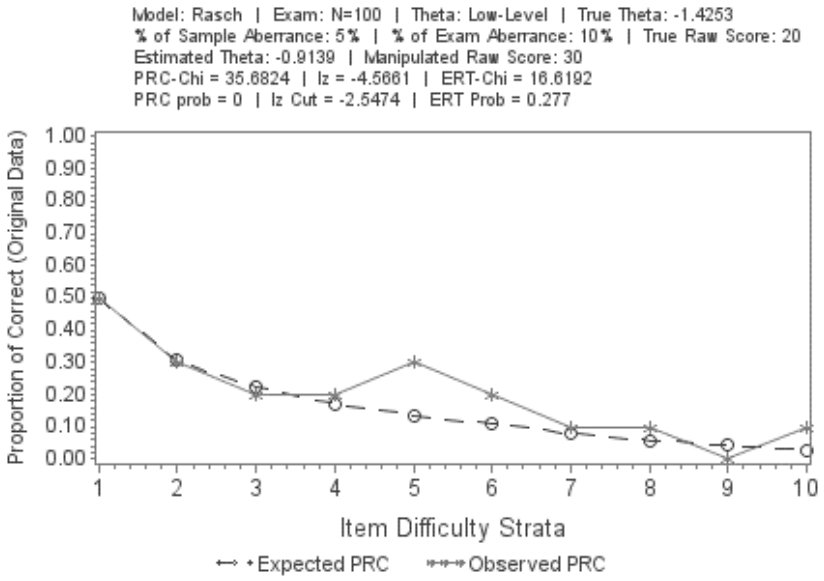

Figure A11. Baseline PRC for condition: Rasch model x long form x low ability.

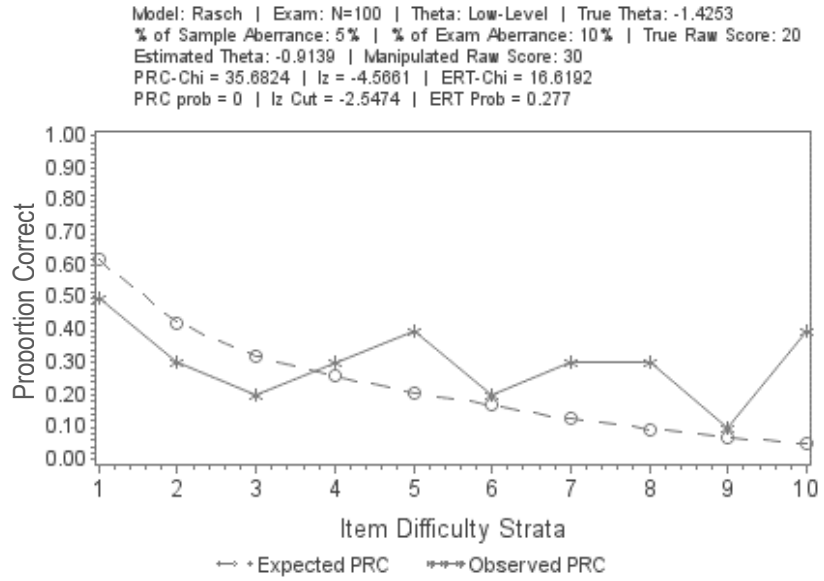

Figure A12. PRC for cheating condition: Rasch model x long form x low ability x 5% sample aberrance x 10% exam aberrance.

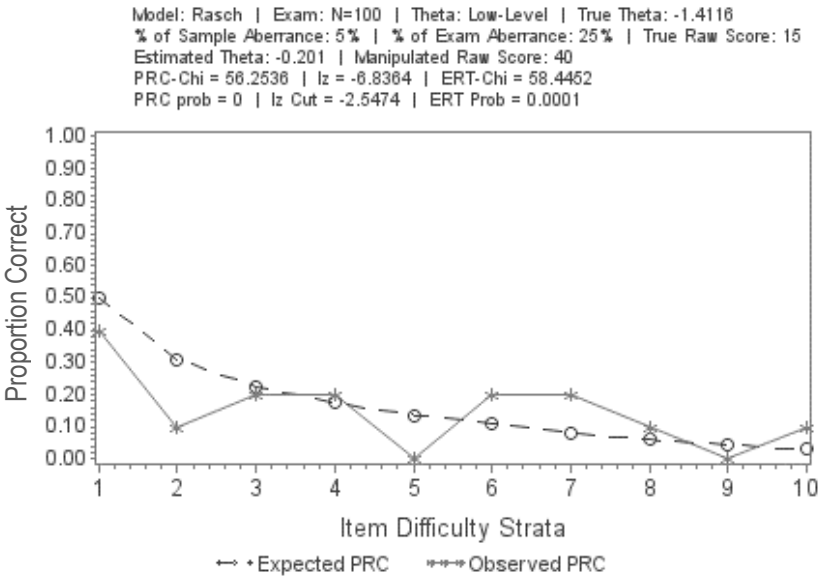

Figure A13. Baseline PRC for condition: Rasch model x long form x low ability.

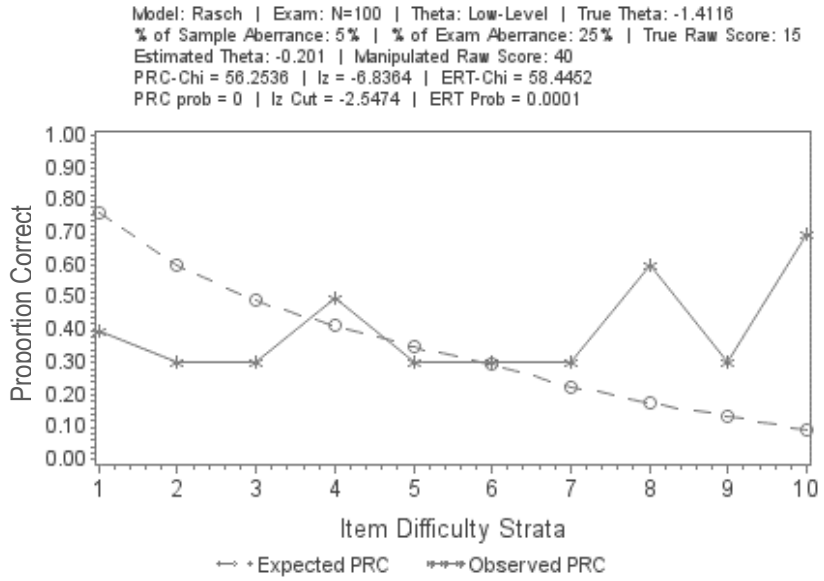

Figure A14. PRC for cheating condition: Rasch model x long form x low ability x 5% sample aberrance x 25% exam aberrance.

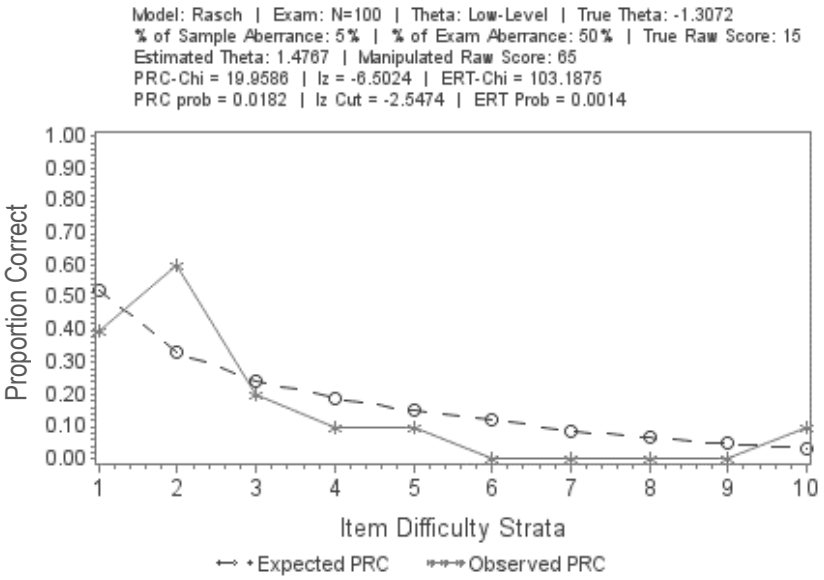

Figure A15. Baseline PRC for condition: Rasch model x long form x low ability.

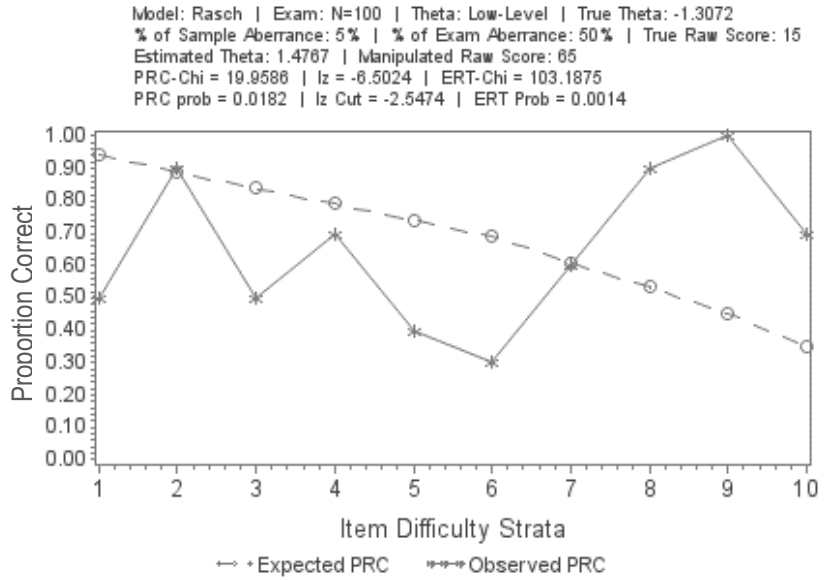

Figure A16. PRC for cheating condition: Rasch model x long form x low ability x 5% sample aberrance x 50% exam aberrance.

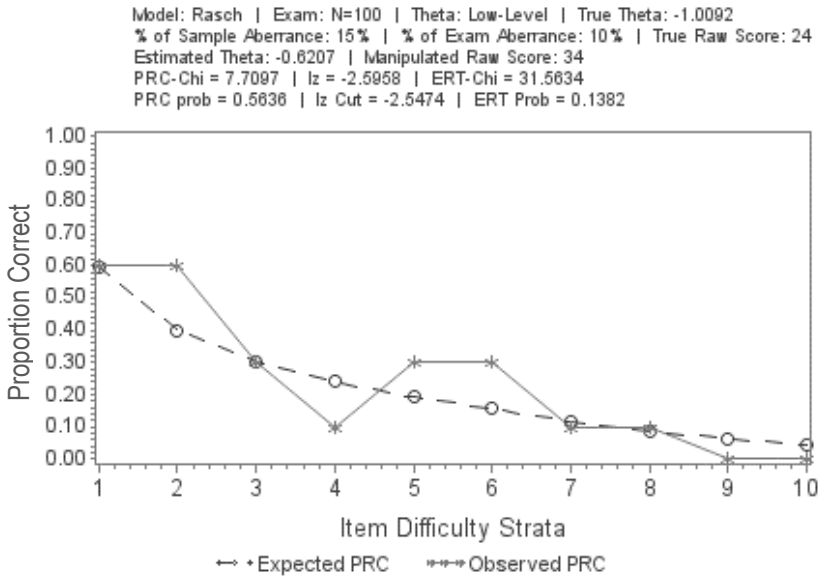

Figure A17. Baseline PRC for condition: Rasch model x long form x low ability.

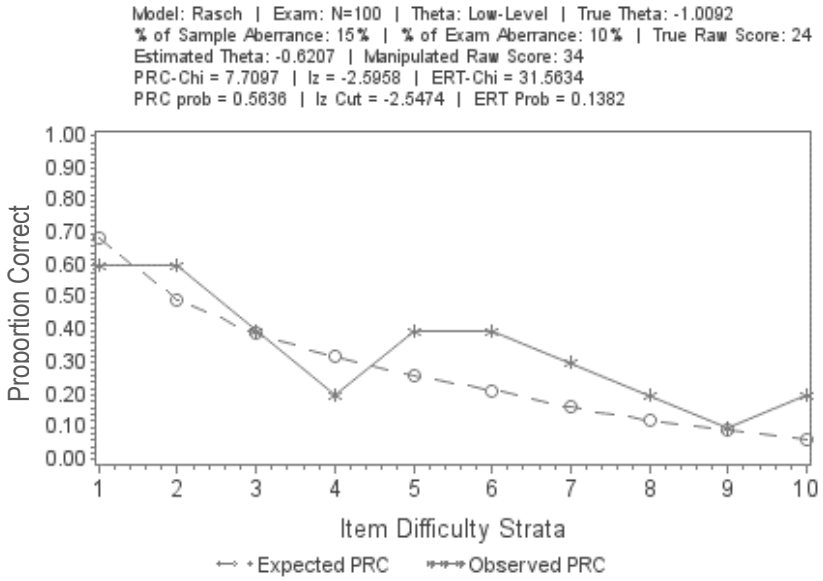

Figure A18. PRC for cheating condition: Rasch model x long form x low ability x 15% sample aberrance x 10% exam aberrance.

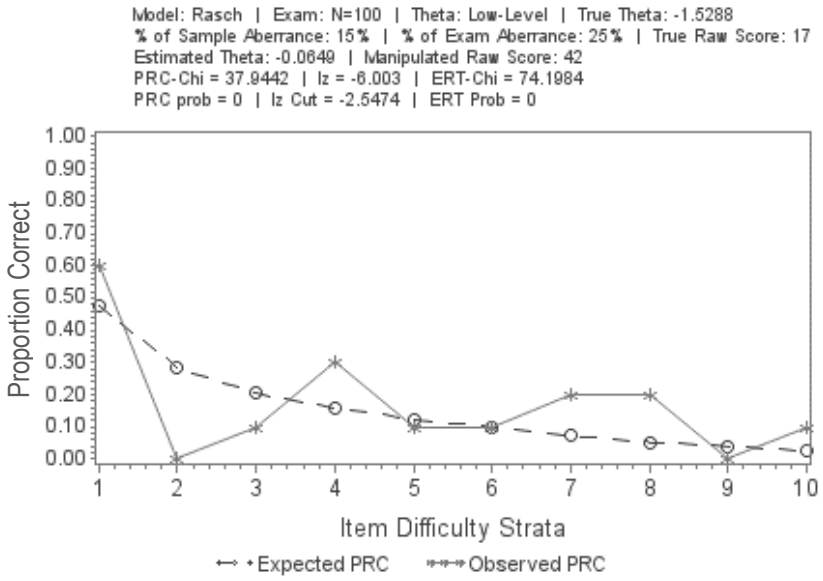

Figure A19. Baseline PRC for condition: Rasch model x long form x low ability.

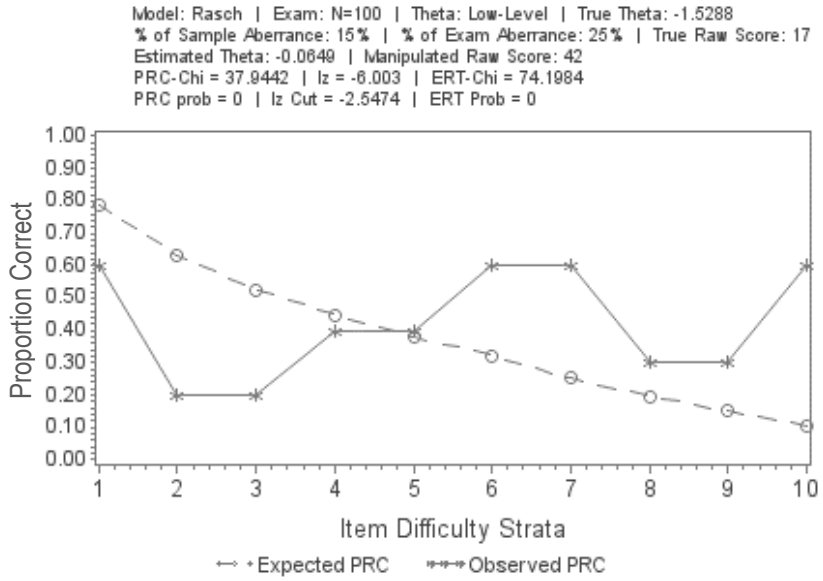

Figure A20. PRC for cheating condition: Rasch model x long form x low ability x 25% exam aberrance x 15% sample aberrance.

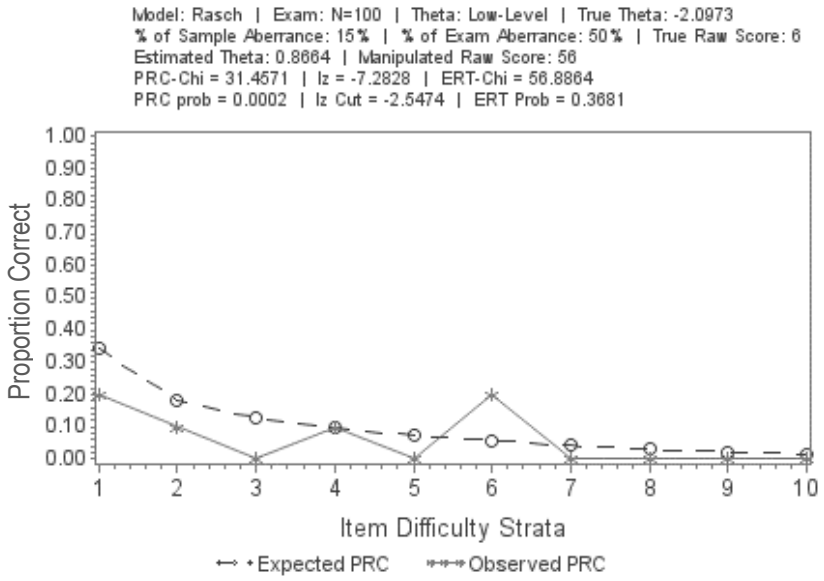

Figure A21. Baseline PRC for condition: Rasch model x long form x low ability.

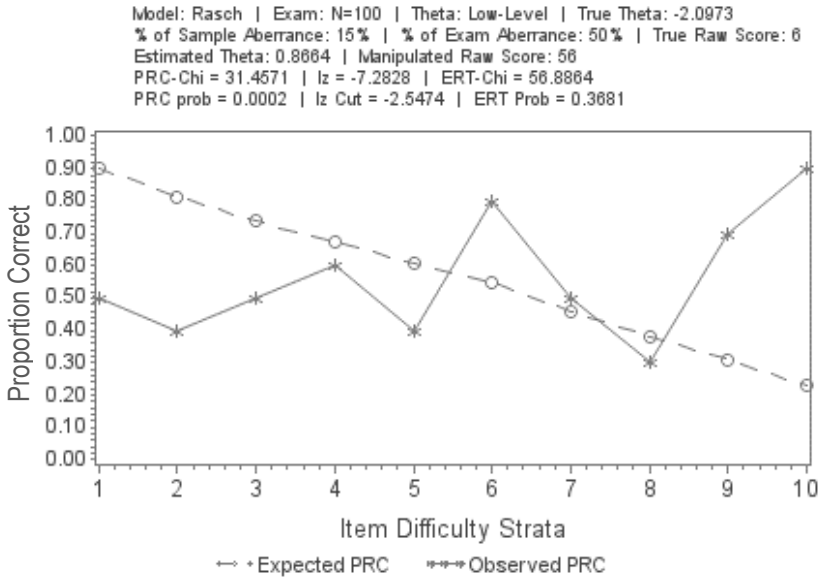

Figure A22. PRC for cheating condition: Rasch model x long form x low ability x 15% sample aberrance x 50% exam aberrance.

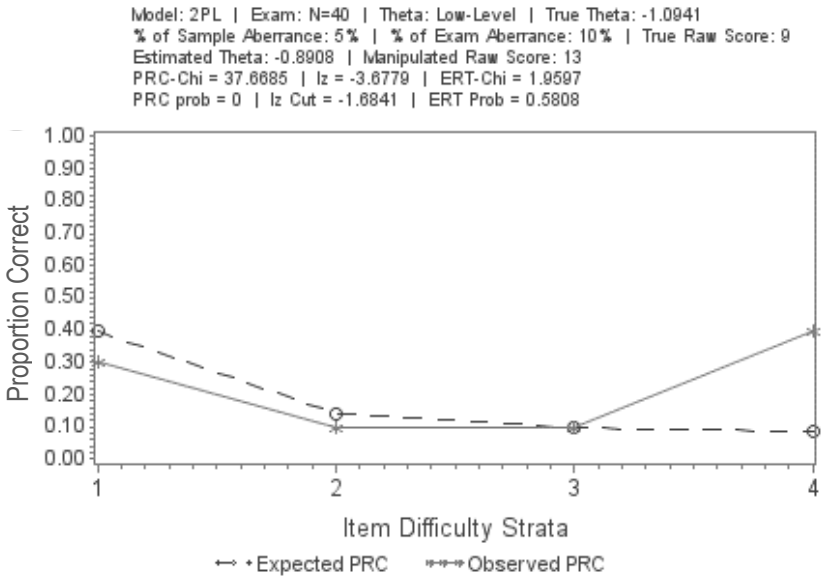

Figure A23. Baseline PRC for condition: 2PL model x short form x low ability.

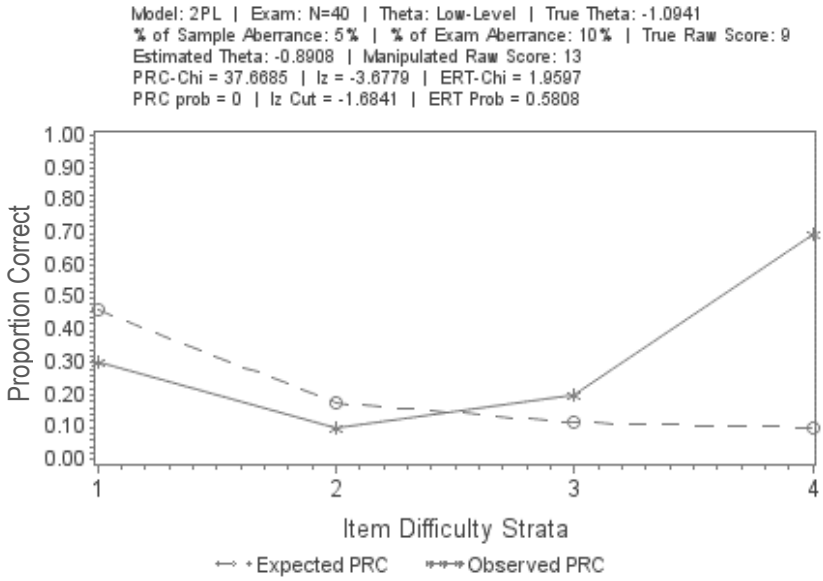

Figure A24. PRC for cheating condition: 2PL model x short form x low ability x 5% sample aberrance x 10% exam aberrance.

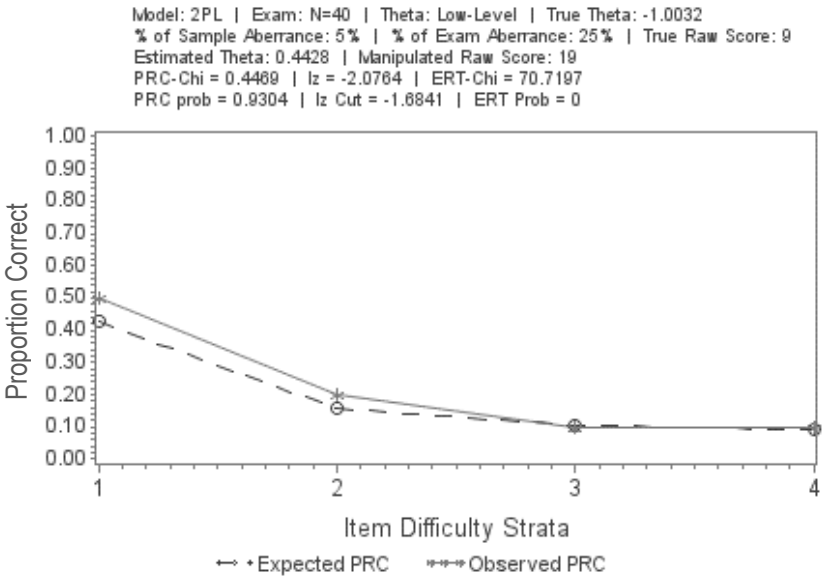

Figure A25. Baseline PRC for condition: 2PL model x short form x low ability.

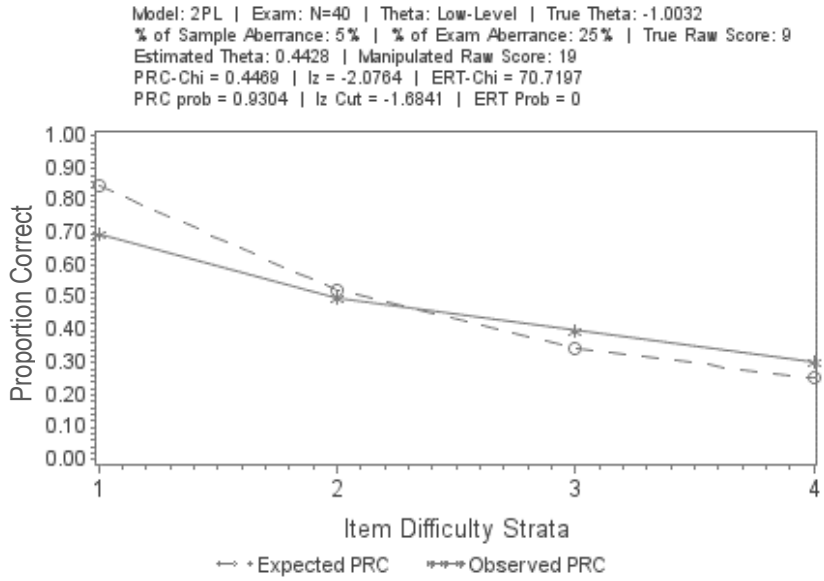

Figure A26. PRC for cheating condition: 2PL model x short form x low ability x 5% sample aberrance x 25% exam aberrance.

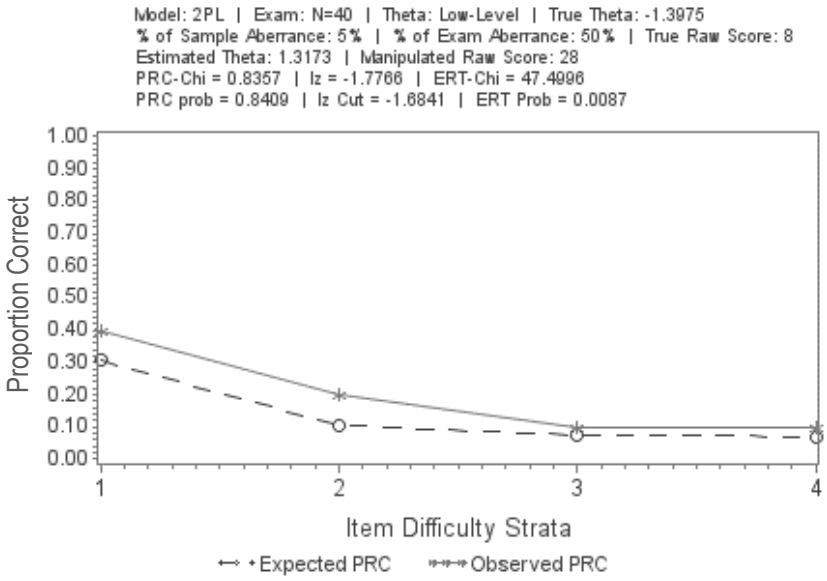

Figure A27. Baseline PRC for condition: 2PL model x short form x low ability.

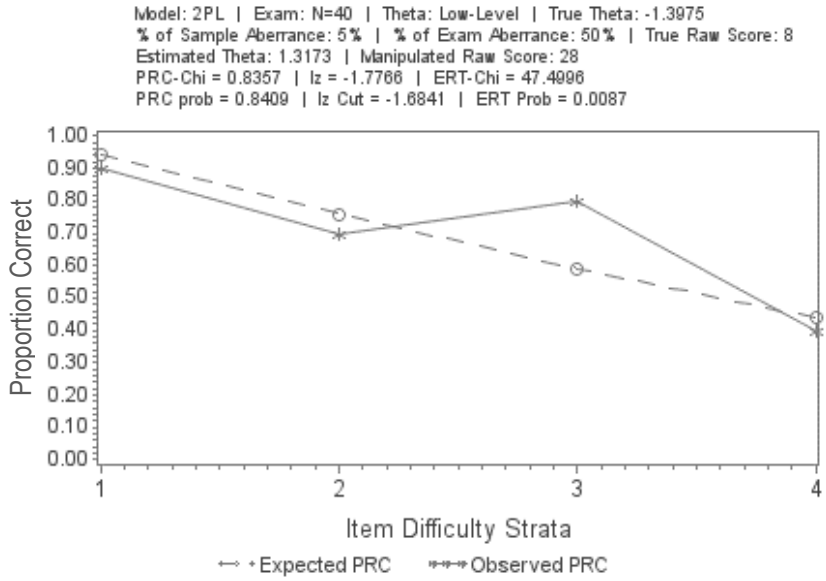

Figure A28. PRC for cheating condition: 2PL model x short form x low ability x 5% sample aberrance x 50% exam aberrance.

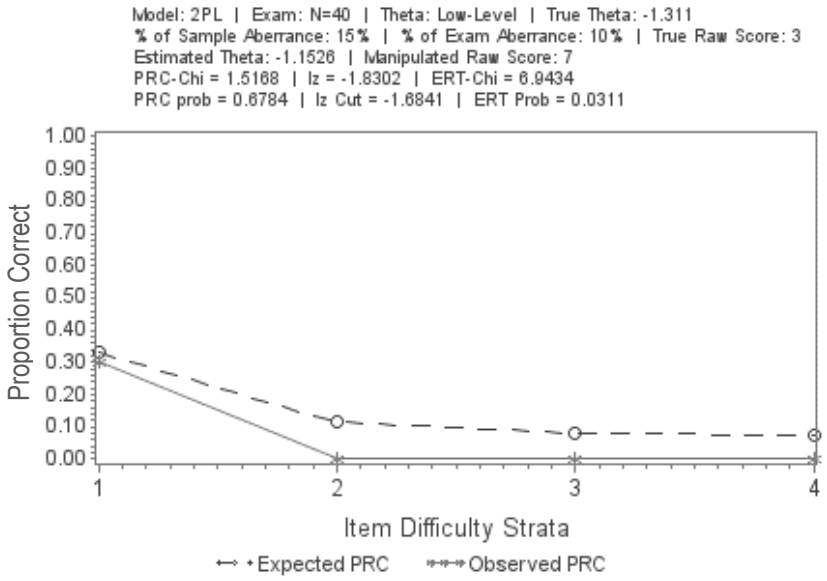

Figure A29. Baseline PRC for condition: 2PL model x short form x low ability.

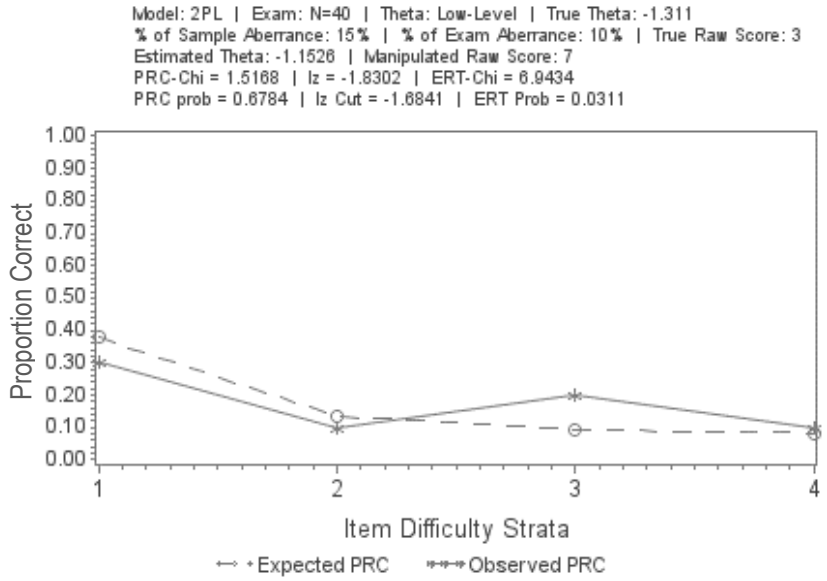

Figure A30. PRC for cheating condition: 2PL model x short form x low ability x 15% sample aberrance x 10% exam aberrance.

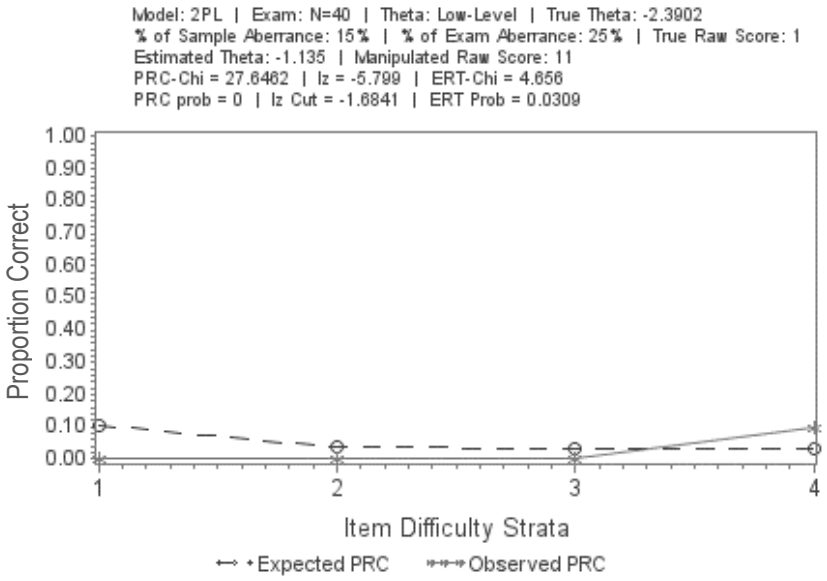

Figure A31. Baseline PRC for condition: 2PL model x short form x low ability.

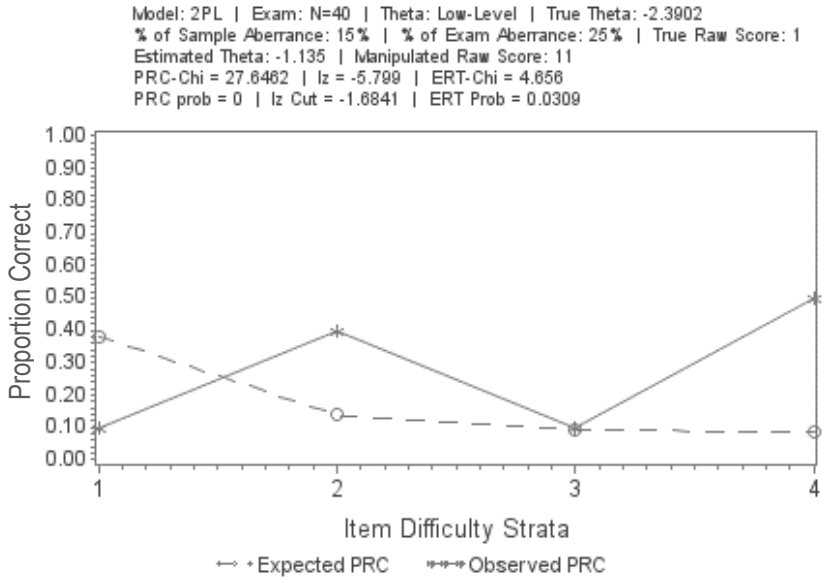

Figure A32. PRC for cheating condition: 2PL model x short form x low ability x 15% sample aberrance x 25% exam aberrance.

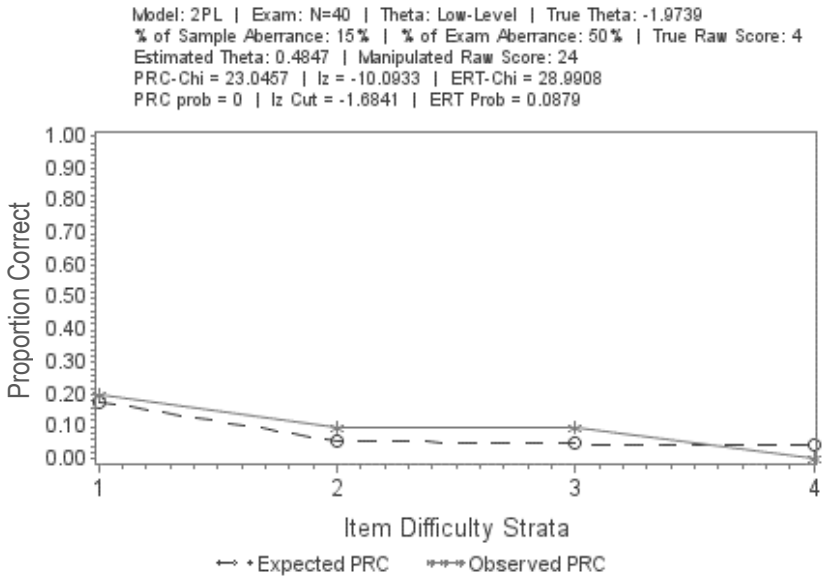

Figure A33. Baseline PRC for condition: 2PL model x short form x low ability.

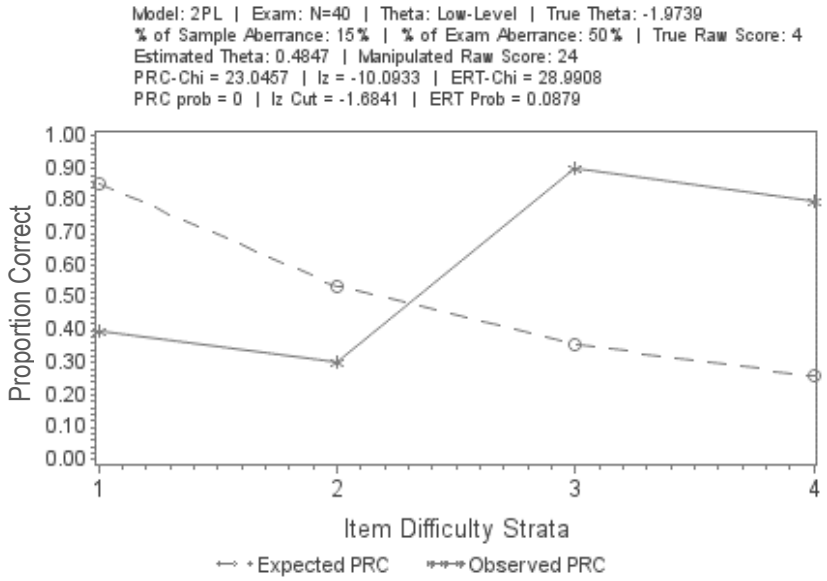

Figure A34. PRC for cheating condition: 2PL model x short form x low ability x 15% sample aberrance x 50% exam aberrance.

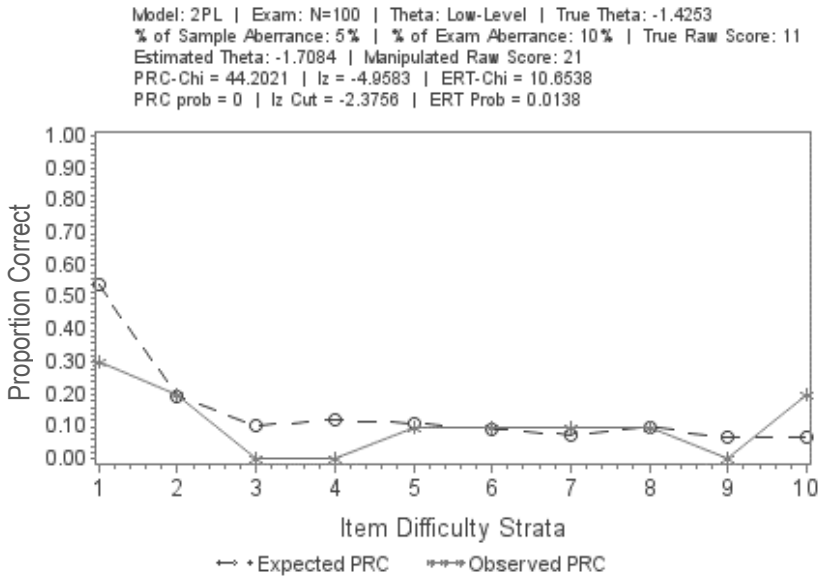

Figure A35. Baseline PRC for condition: 2PL model x long form x low ability.

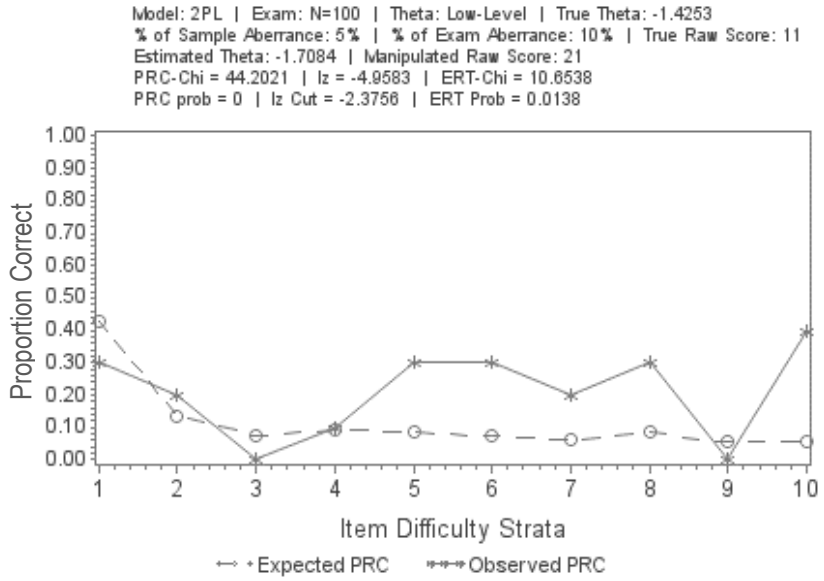

Figure A36. PRC for cheating condition: 2PL model x long form x low ability x 5% sample aberrance x 10% exam aberrance.

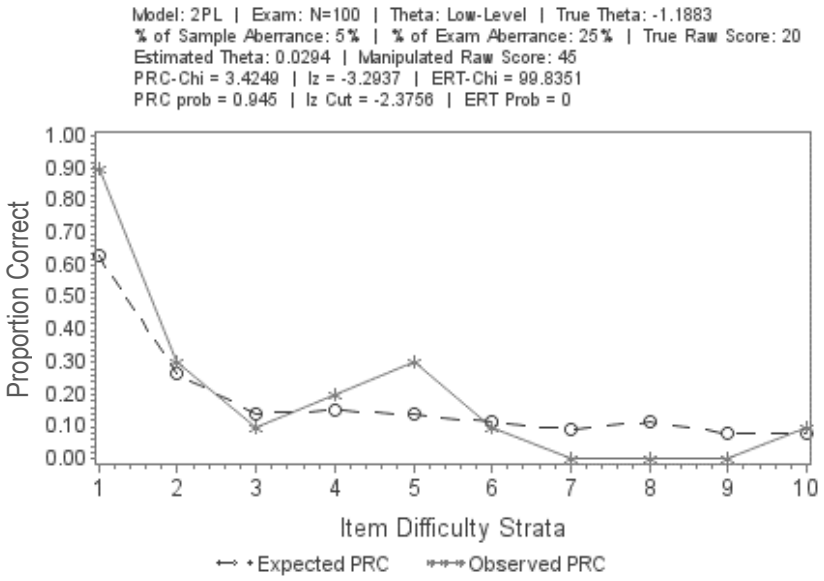

Figure A37. Baseline PRC for condition: 2PL model x long form x low ability.

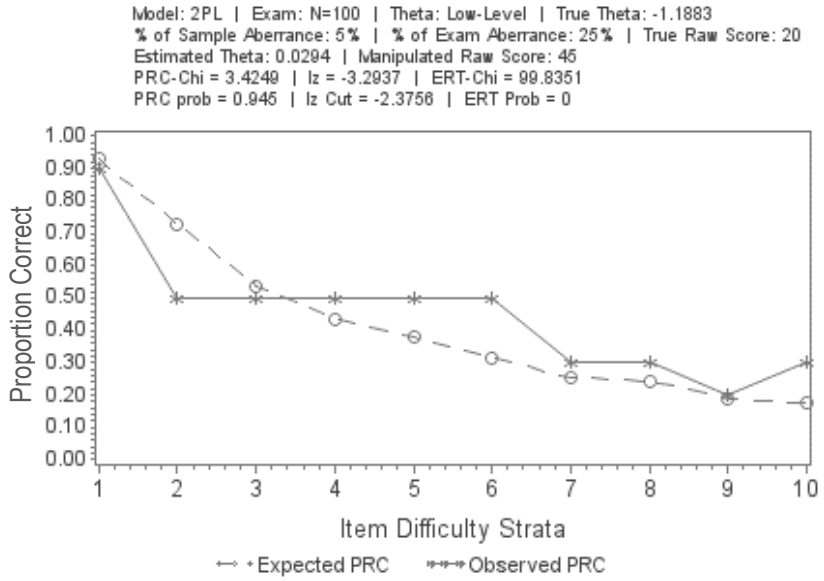

Figure A38. PRC for cheating condition: 2PL model x long form x low ability x 5% sample aberrance x 25% exam aberrance.

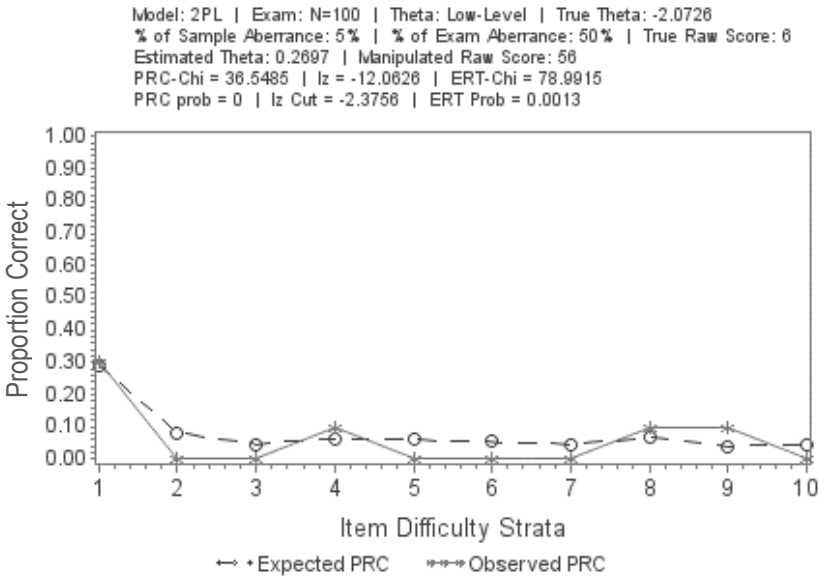

Figure A39. Baseline PRC for condition: 2PL model x long form x low ability.

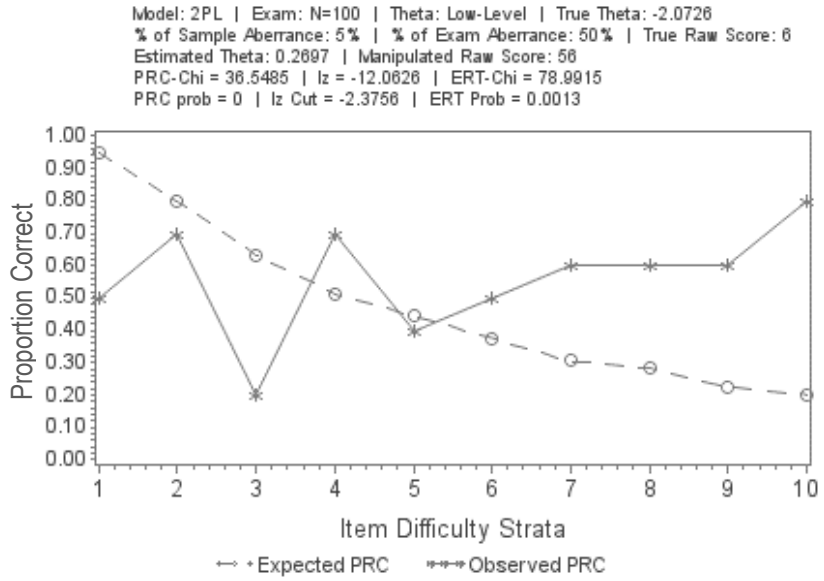

Figure A40. PRC for cheating condition: 2PL model x long form x low ability x 5% sample aberrance x 50% exam aberrance.

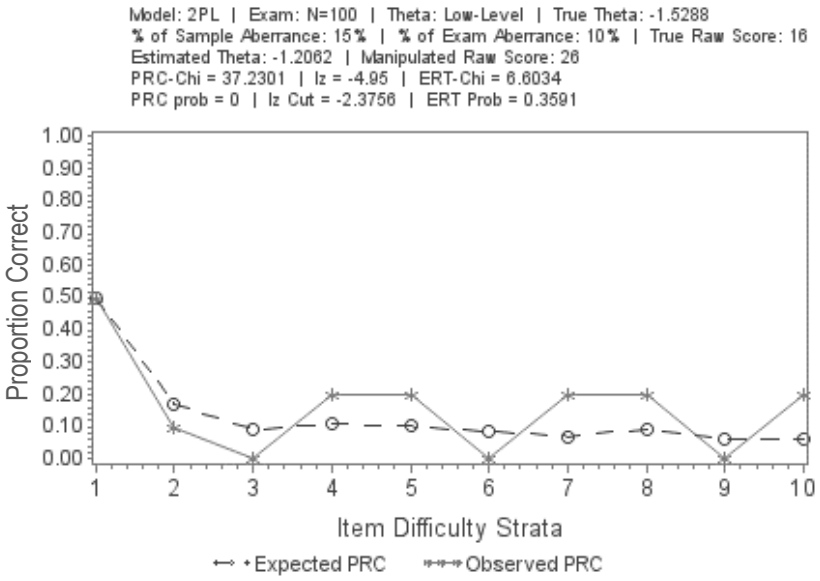

Figure A41. Baseline PRC for condition: 2PL model x long form x low ability.

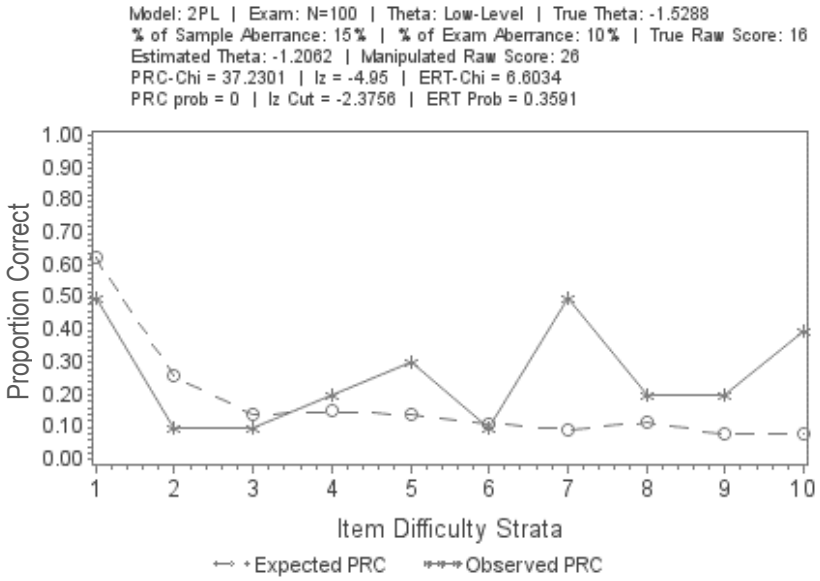

Figure A42. PRC for cheating condition: 2PL model x long form x low ability x 15% sample aberrance x 10% exam aberrance.

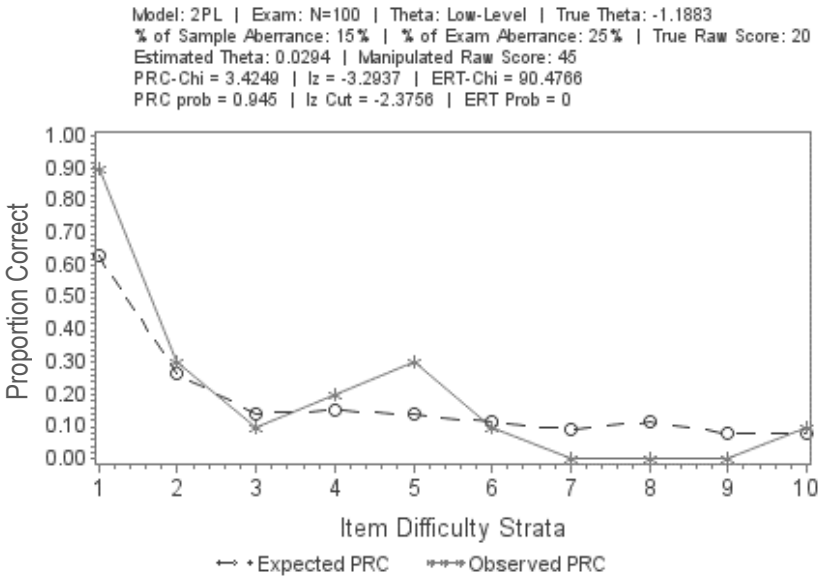

Figure A43. Baseline PRC for condition: 2PL model x long form x low ability.

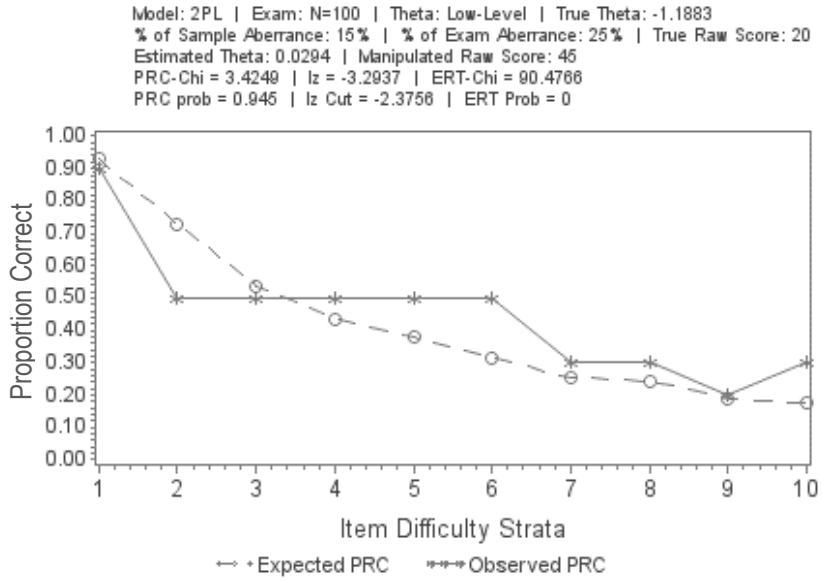

Figure A44. PRC for cheating condition: 2PL model x long form x low ability x 15% sample aberrance x 25% exam aberrance.

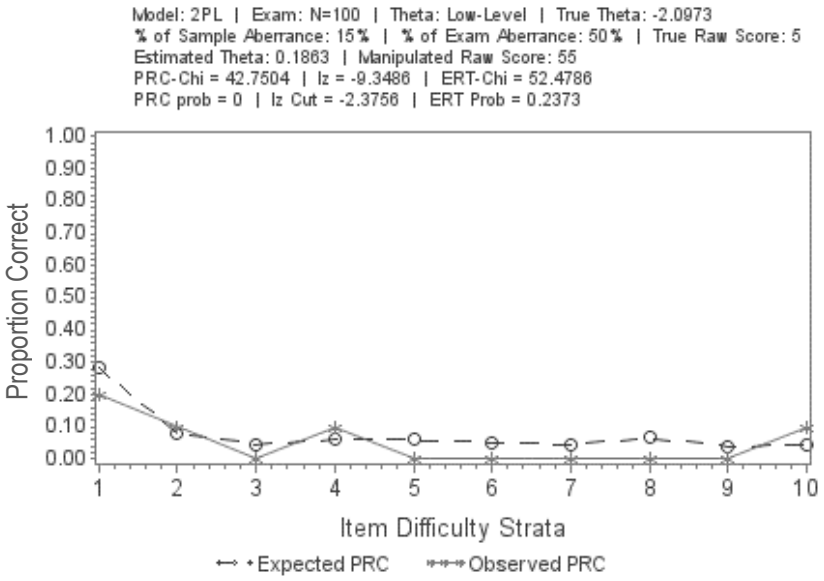

Figure A45. Baseline PRC for condition: 2PL model x long form x low ability.

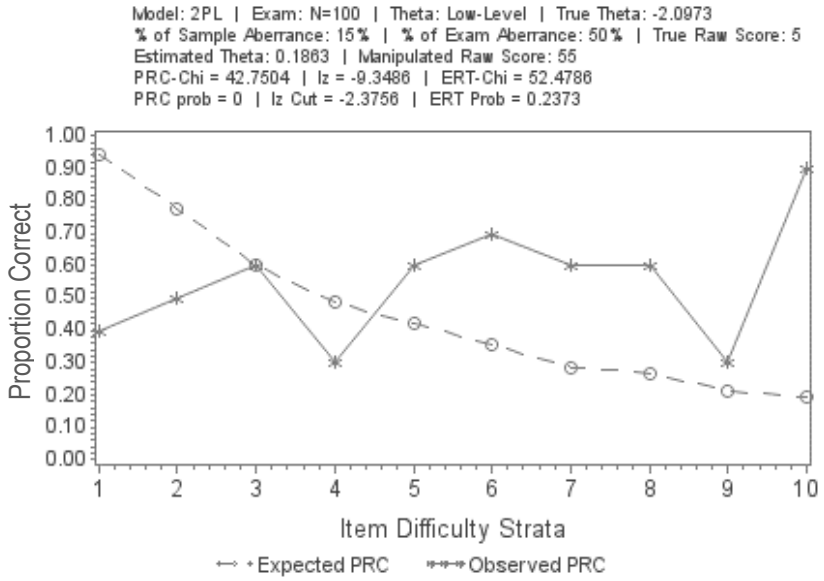

Figure A46. PRC for cheating condition: 2PL model x long form x low ability x 15% sample aberrance x 50% exam aberrance.

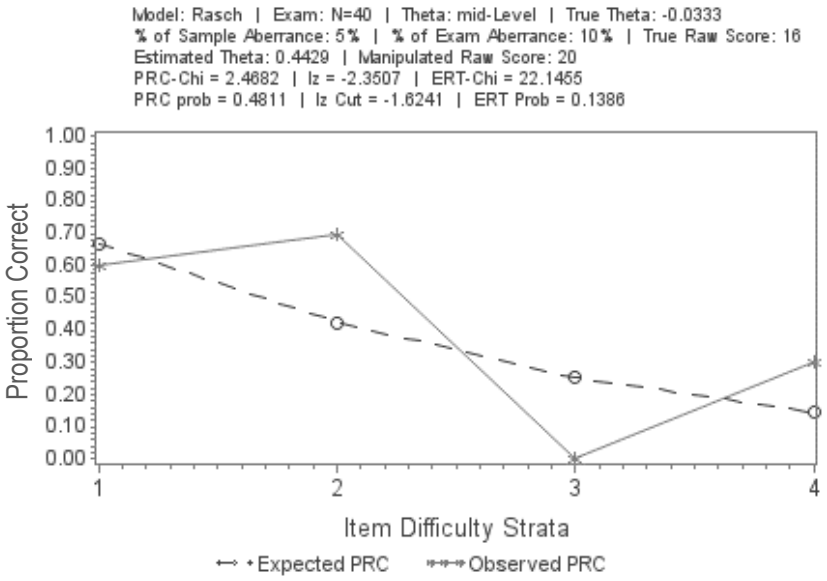

Figure A47. Baseline PRC for condition: Rasch model x short form x mid ability.

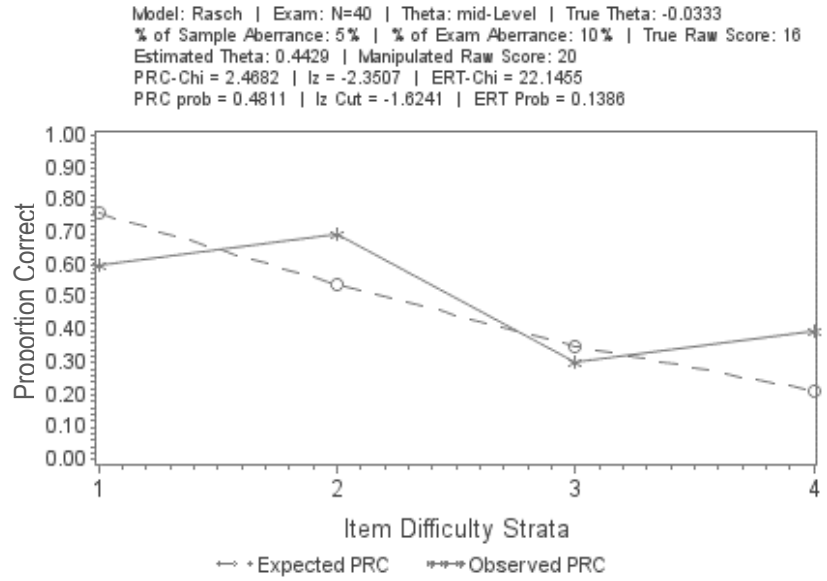

Figure A48. PRC for cheating condition: Rasch model x short form x mid ability x 5% sample aberrance x 10% exam aberrance.

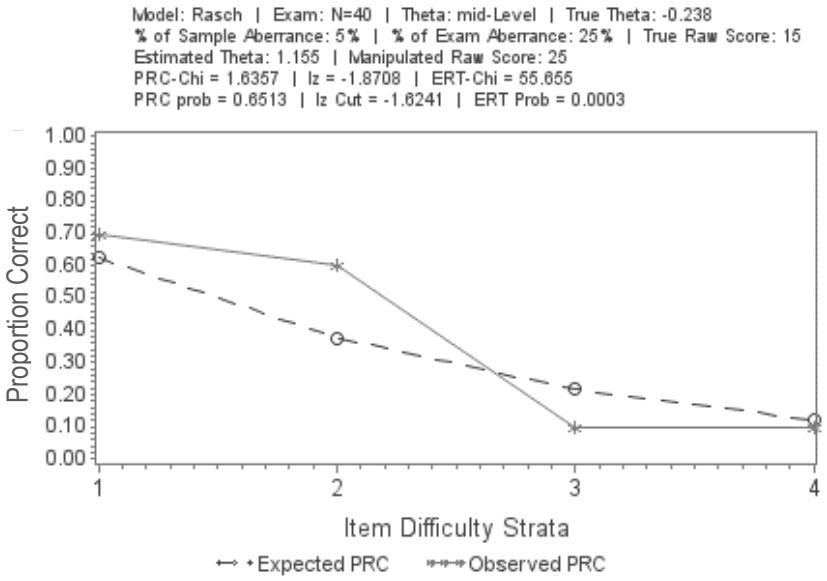

Figure A49. Baseline PRC for condition: Rasch model x short form x mid ability.

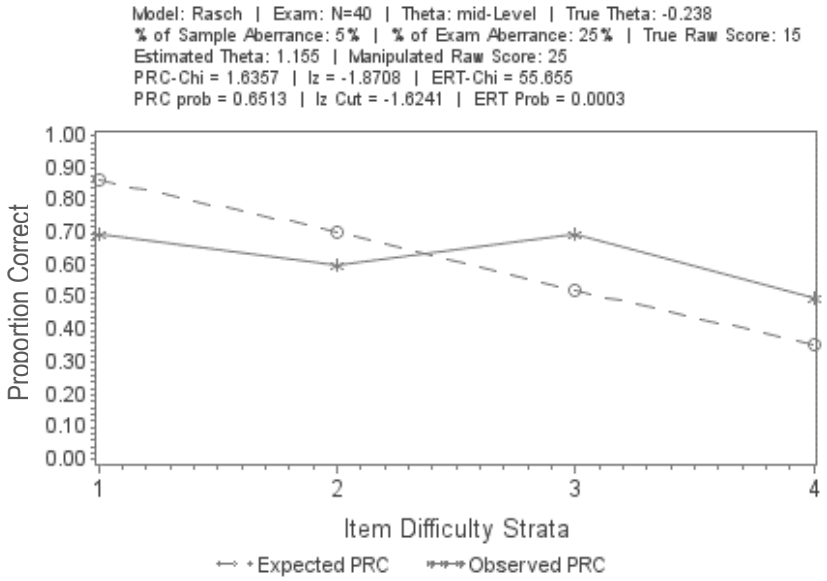

Figure A50. PRC for cheating condition: Rasch model x short form x mid ability x 5% sample aberrance x 25% exam aberrance.

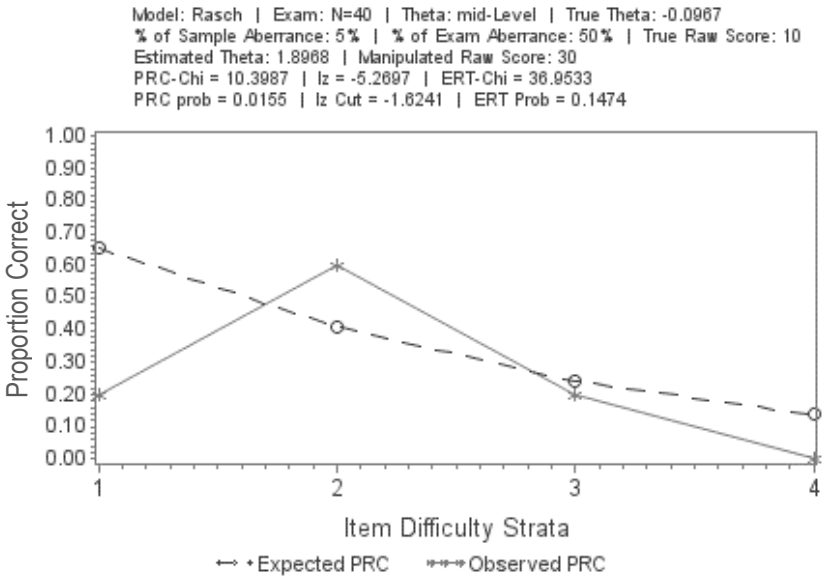

Figure A51. Baseline PRC for condition: Rasch model x short form x mid ability.

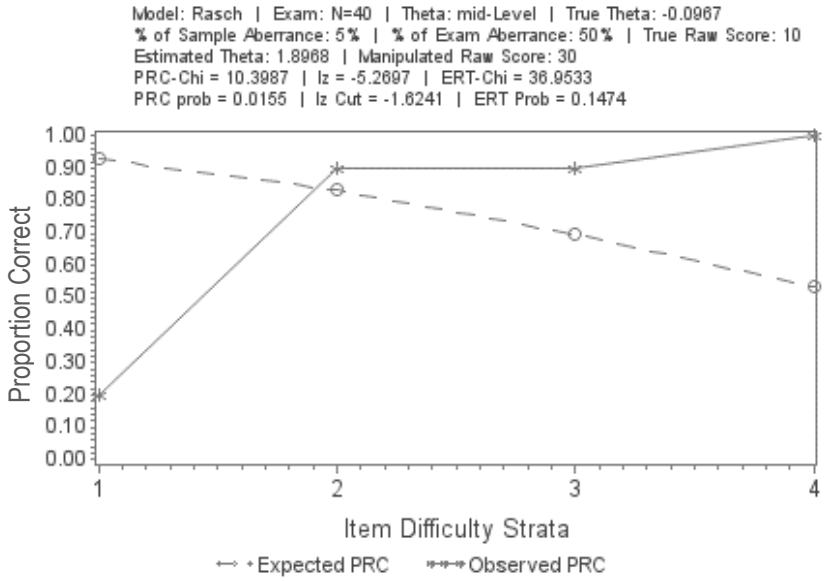

Figure A52. PRC for cheating condition: Rasch model x short form x mid ability x 5% sample aberrance x 50% exam aberrance.

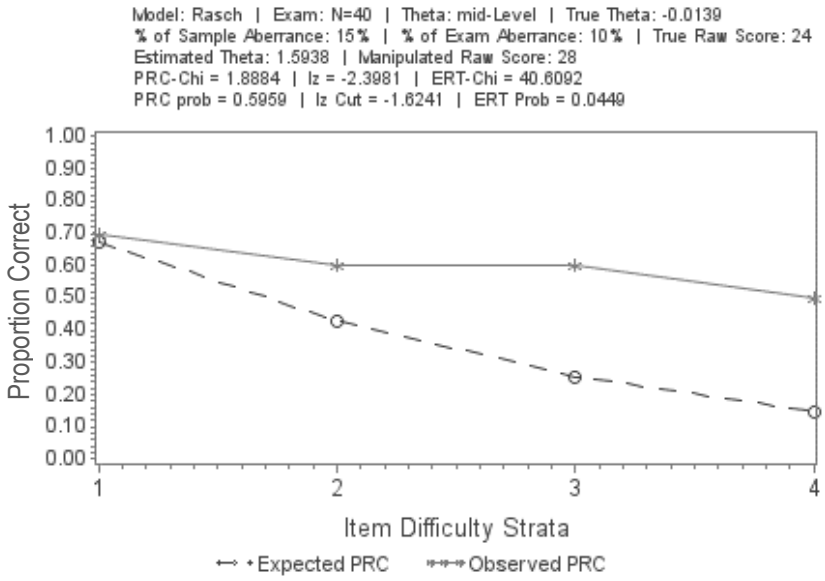

Figure A53. Baseline PRC for condition: Rasch model x short form x mid ability.

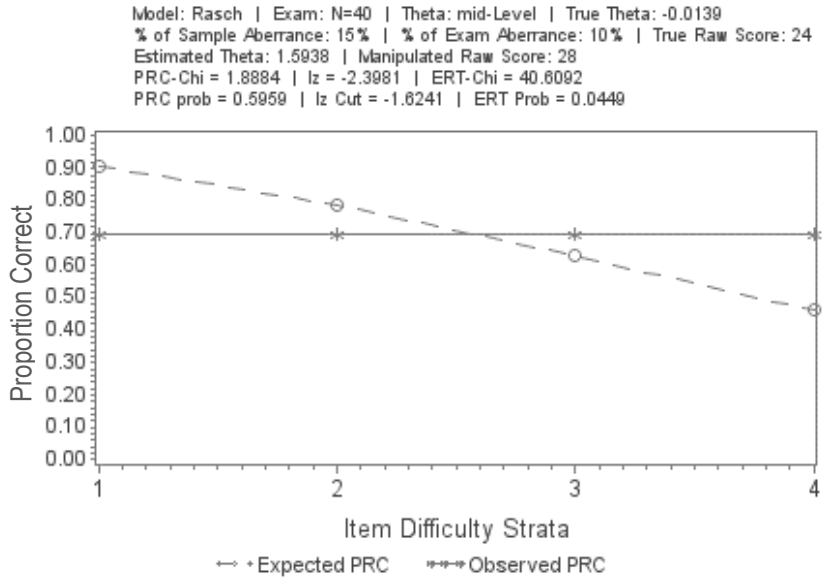

Figure A54. PRC for cheating condition: Rasch model x short form x mid ability x 15% sample aberrance x 10% exam aberrance.

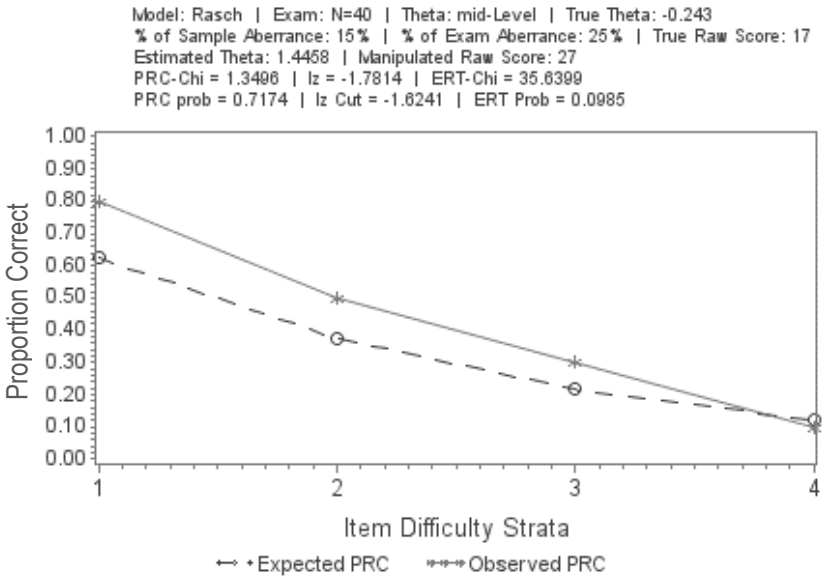

Figure A55. Baseline PRC for condition: Rasch model x short form x mid ability.

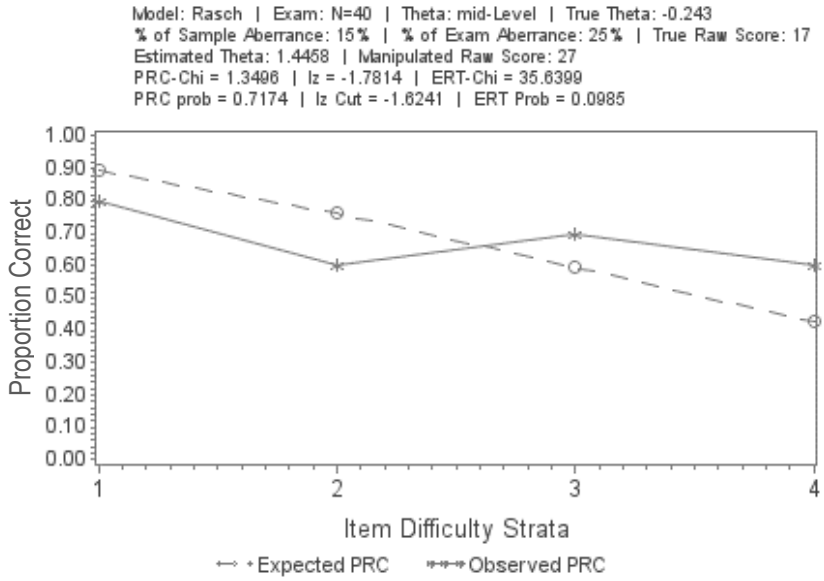

Figure A56. PRC for cheating condition: Rasch model x short form x mid ability x 15% sample aberrance x 25% exam aberrance.

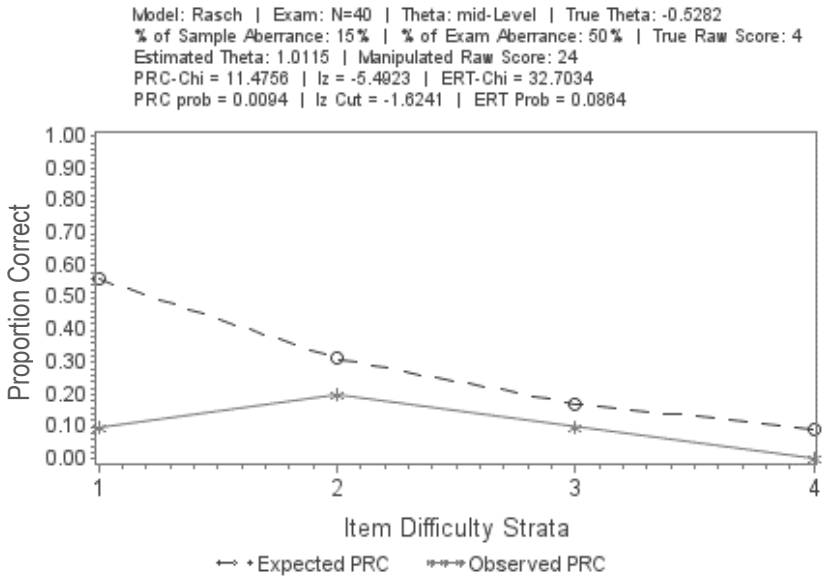

Figure A57. Baseline PRC for condition: Rasch model x short form x mid ability.

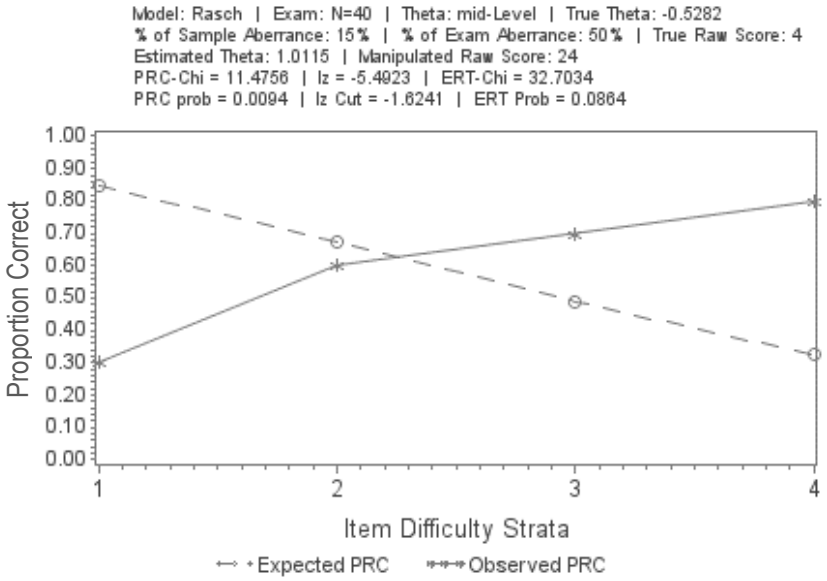

Figure A58. PRC for cheating condition: Rasch model x short form x mid ability x 15% sample aberrance x 50% exam aberrance.

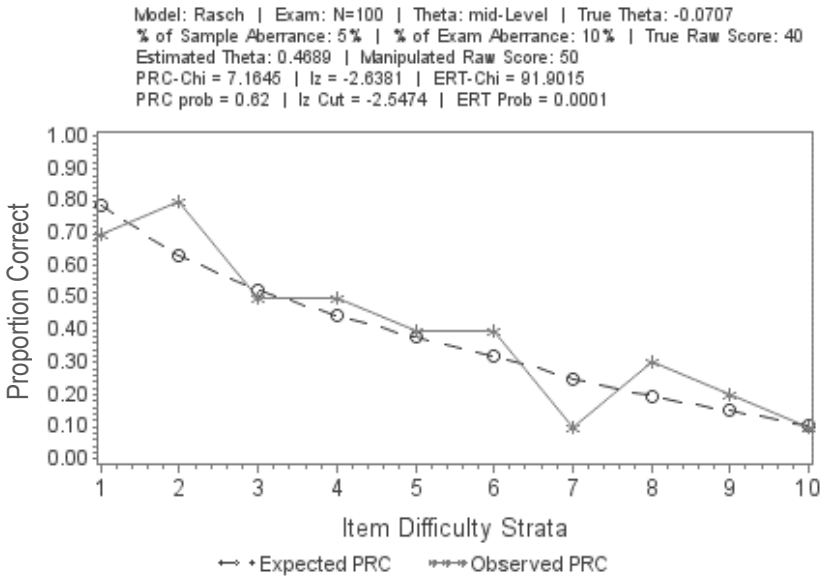

Figure A59. Baseline PRC for condition: Rasch model x long form x mid ability.

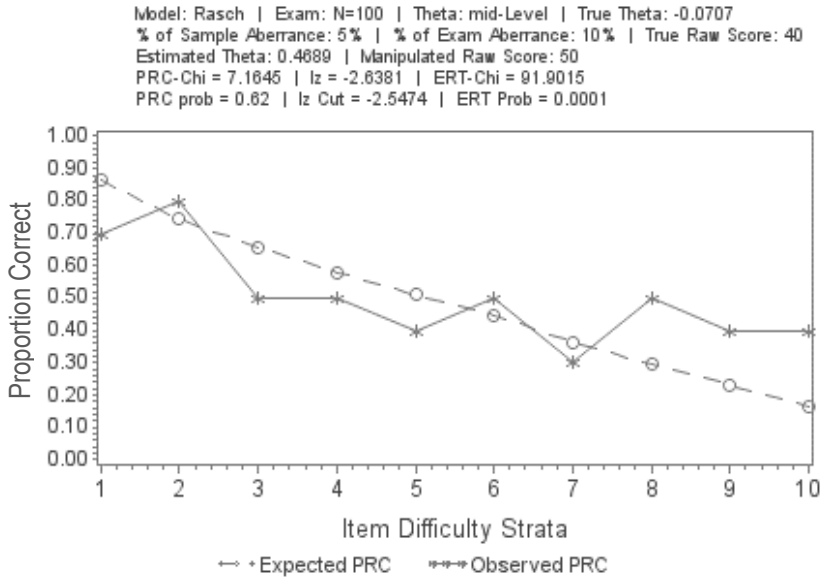

Figure A60. PRC for cheating condition: Rasch model x long form x mid ability x 5% sample aberrance x 10% exam aberrance.

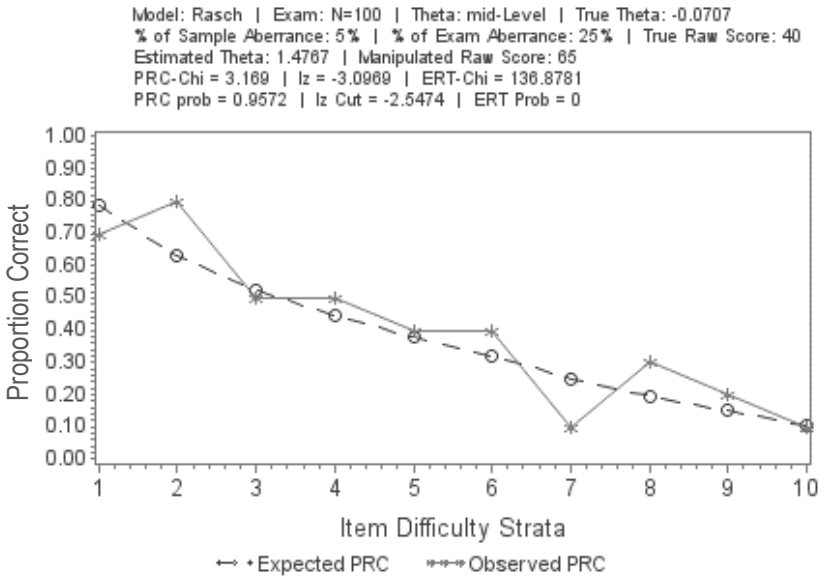

Figure A61. Baseline PRC for condition: Rasch model x long form x mid ability.

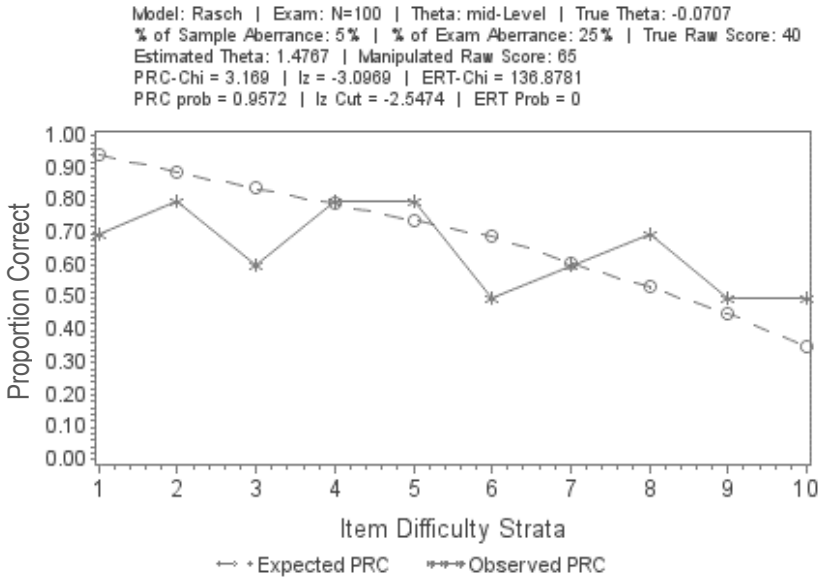

Figure A62. PRC for cheating condition: Rasch model x long form x mid ability x 5% sample aberrance x 25% exam aberrance.

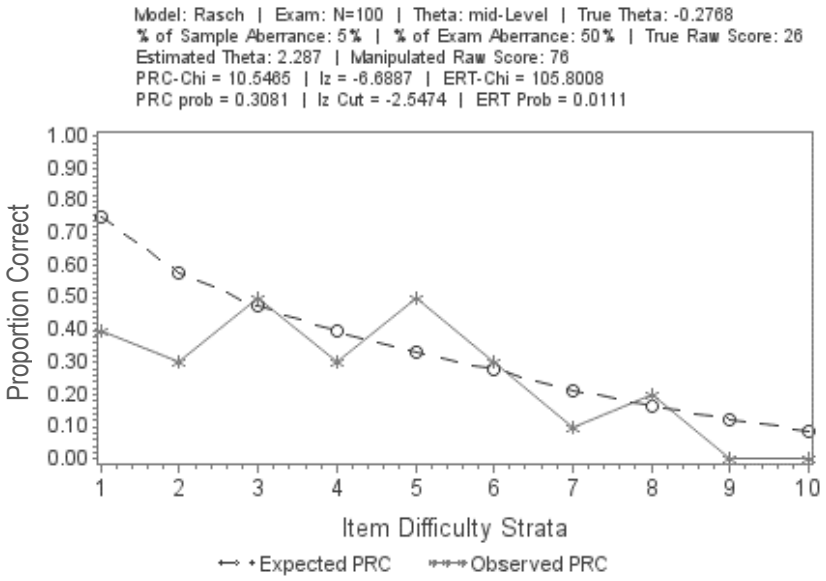

Figure A63. Baseline PRC for condition: Rasch model x long form x mid ability.

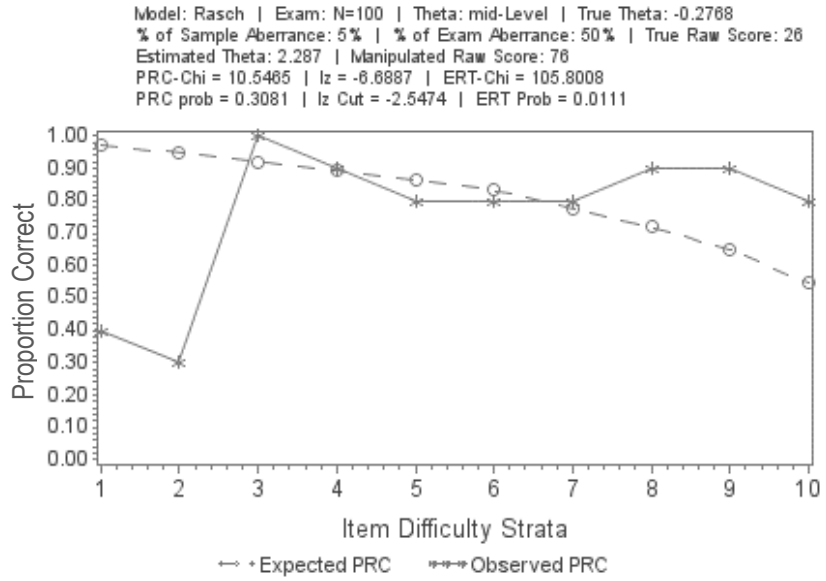

Figure A64. PRC for cheating condition: Rasch model x long form x mid ability x 5% sample aberrance x 50% exam aberrance.

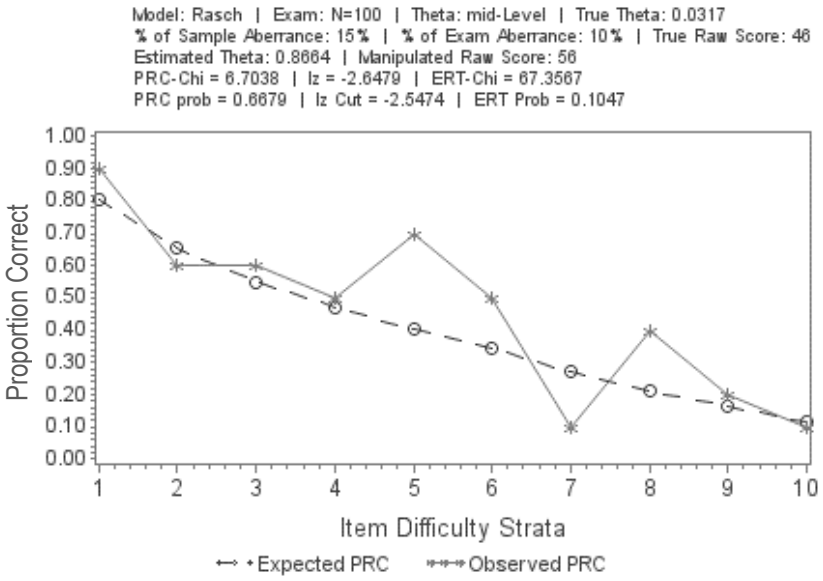

Figure A65. Baseline PRC for condition: Rasch model x long form x mid ability.

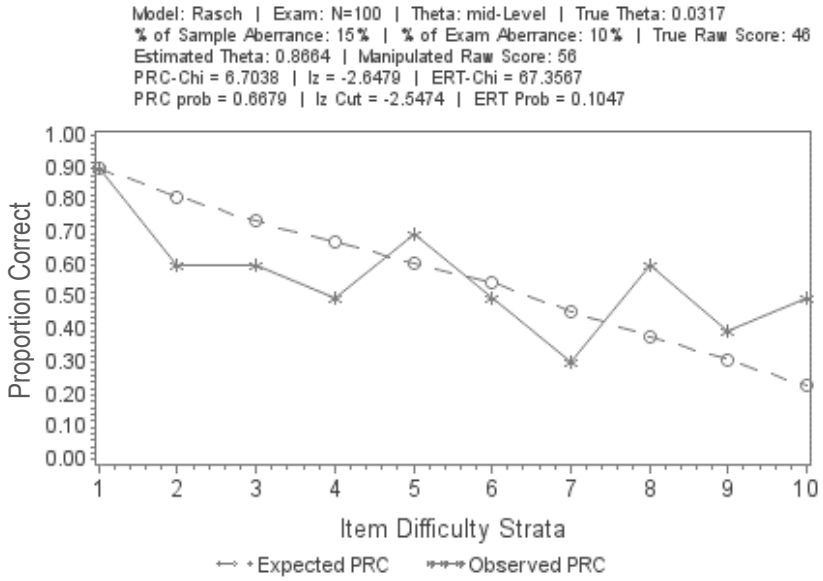

Figure A66. PRC for cheating condition: Rasch model x long form x mid ability x 15% sample aberrance x 10% exam aberrance.

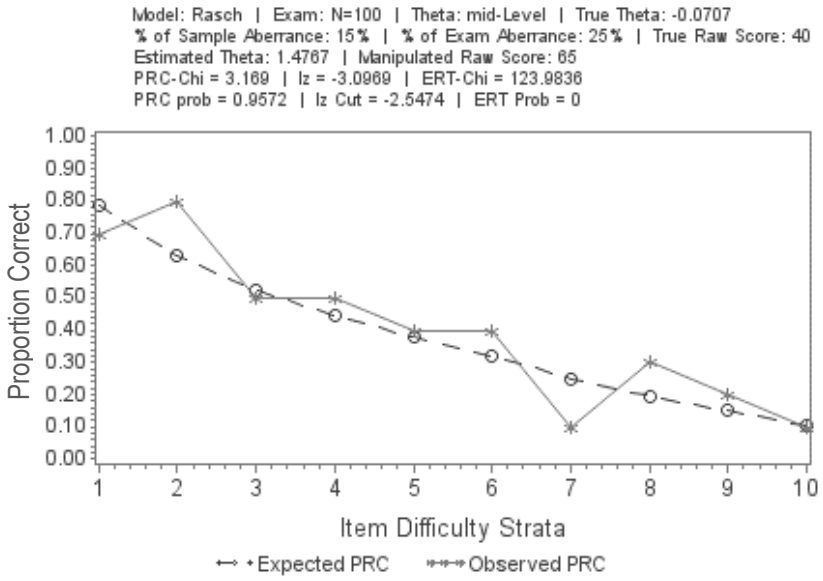

Figure A67. Baseline PRC for condition: Rasch model x long form x mid ability.

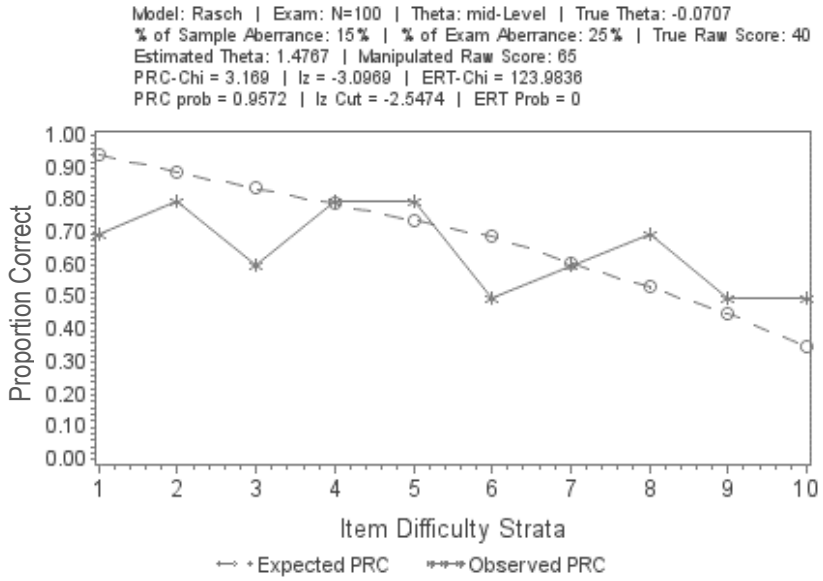

Figure A68. PRC for cheating condition: Rasch model x long form x mid ability x 15% sample aberrance x 25% exam aberrance.

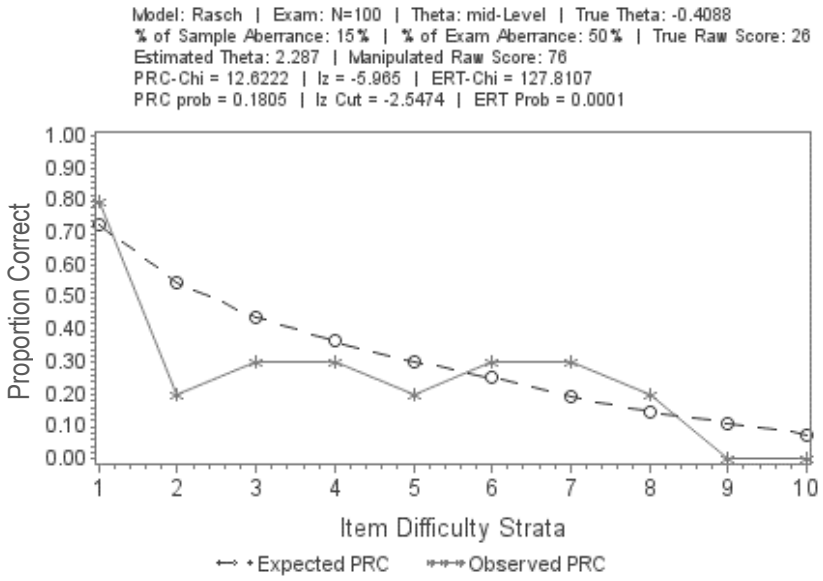

Figure A69. Baseline PRC for condition: Rasch model x long form x mid ability.

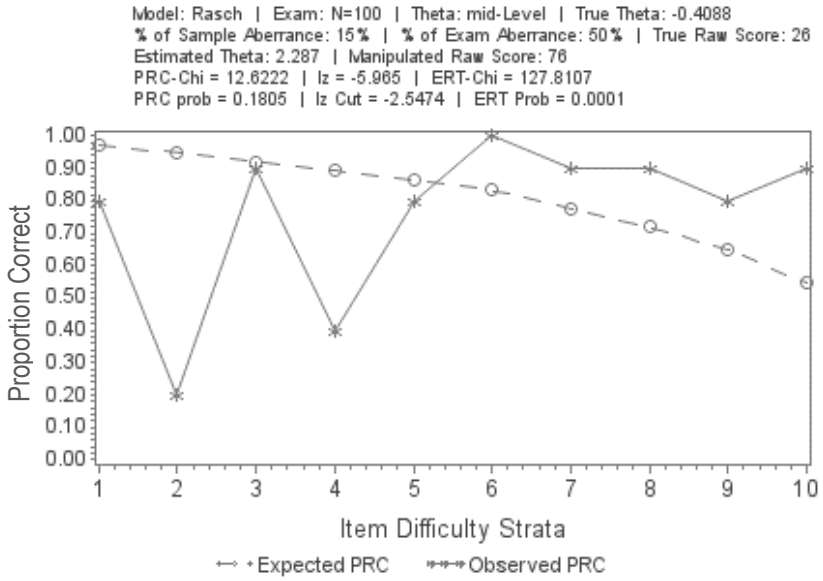

Figure A70. PRC for cheating condition: Rasch model x long form x mid ability x 15% sample aberrance x 50% exam aberrance.

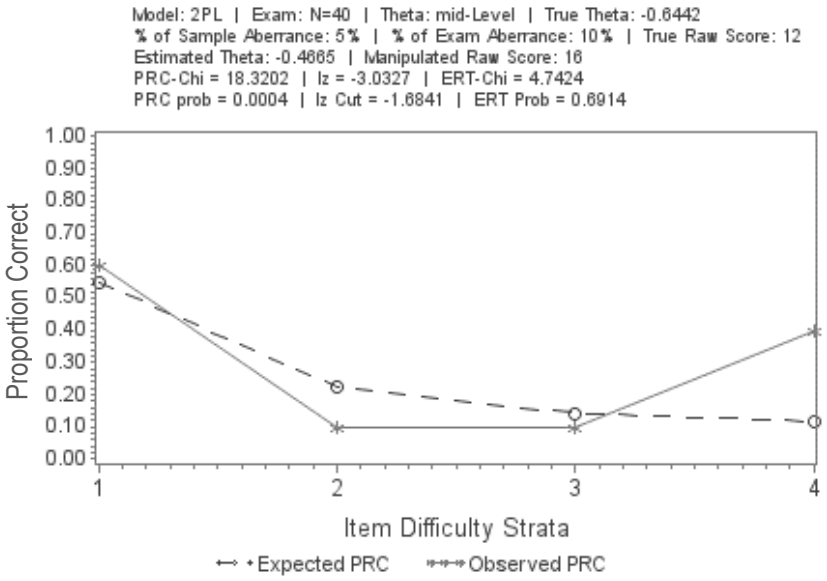

Figure A71. Baseline PRC for condition: 2PL model x short form x mid ability.

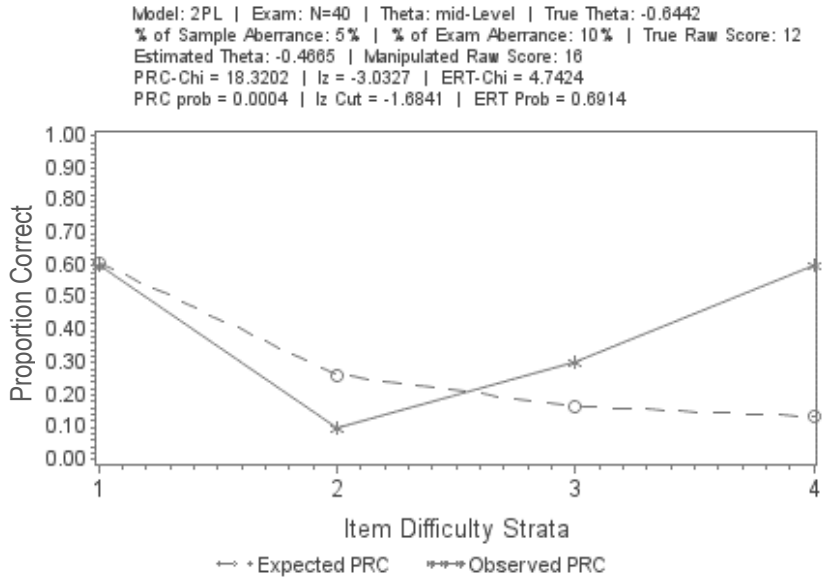

Figure A72. PRC for cheating condition: 2PL model x short form x mid ability x 5% sample aberrance x 10% exam aberrance.

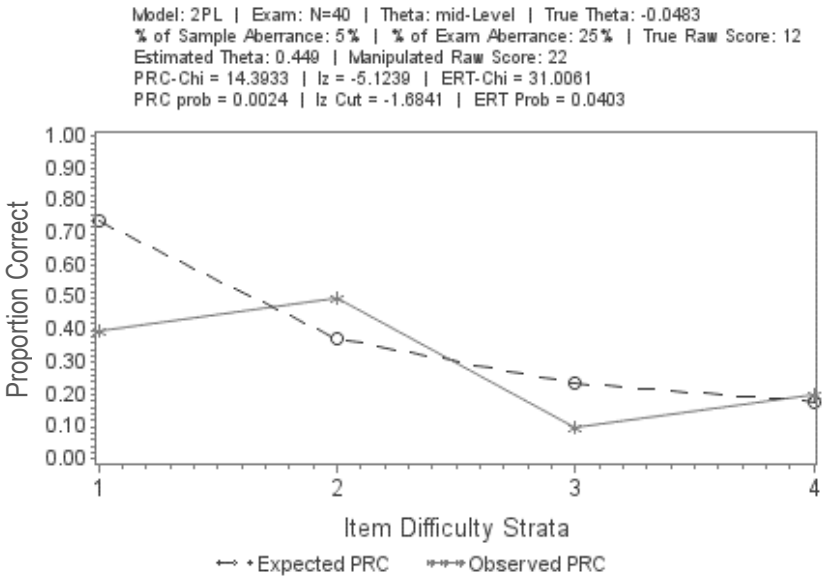

Figure A73. Baseline PRC for condition: 2PL model x short form x mid ability.

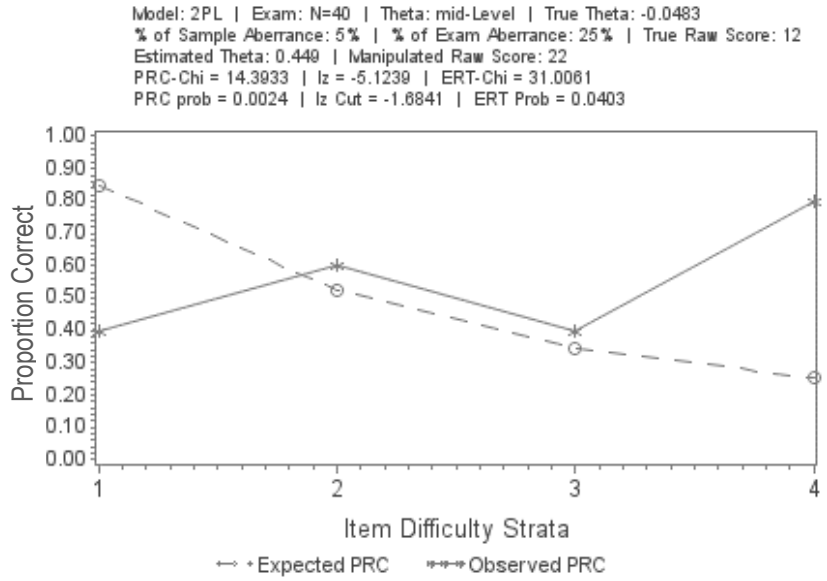

Figure A74. PRC for cheating condition: 2PL model x short form x mid ability x 5% sample aberrance x 25% exam aberrance.

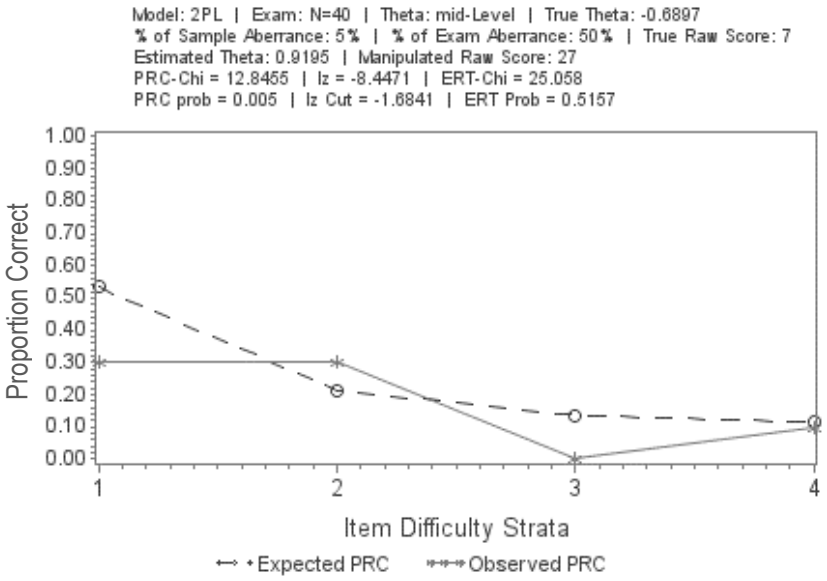

Figure A75. Baseline PRC for condition: 2PL model x short form x mid ability.

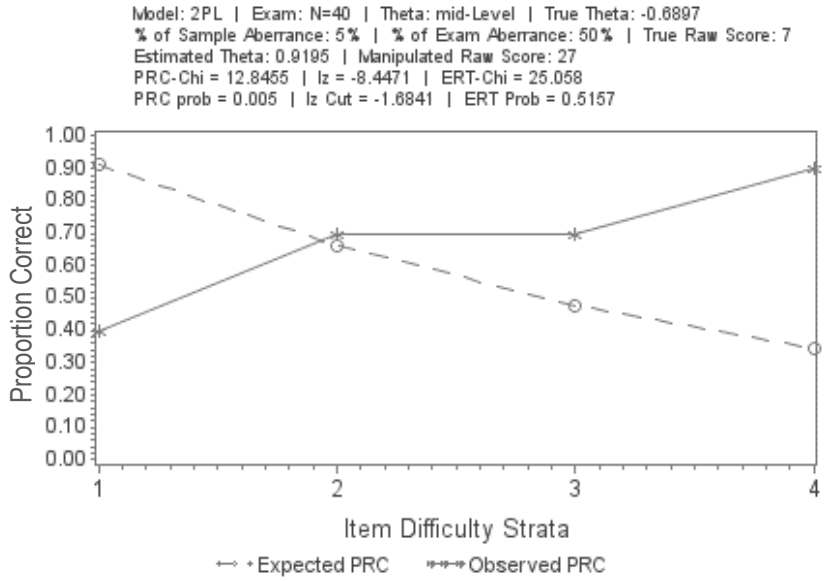

Figure A76. PRC for cheating condition: 2PL model x short form x mid ability x 5% sample aberrance x 50% exam aberrance.

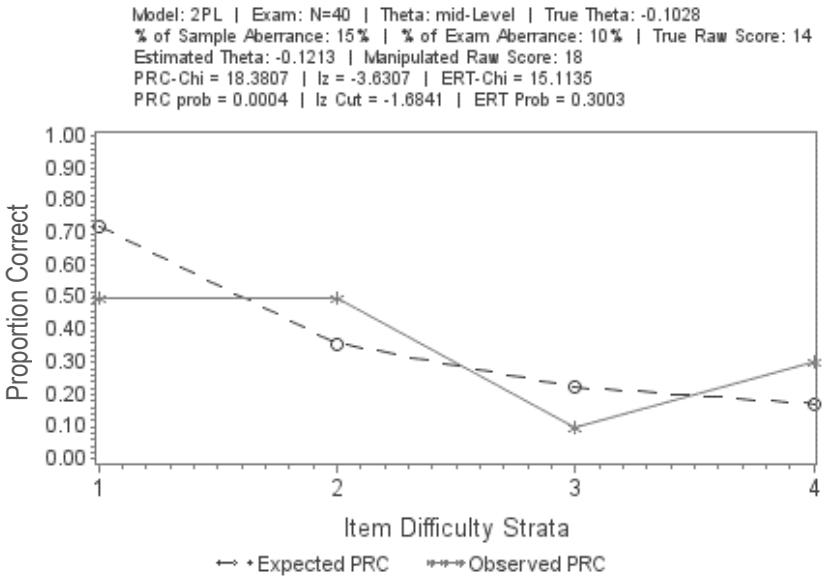

Figure A77. Baseline PRC for condition: 2PL model x short form x mid ability.

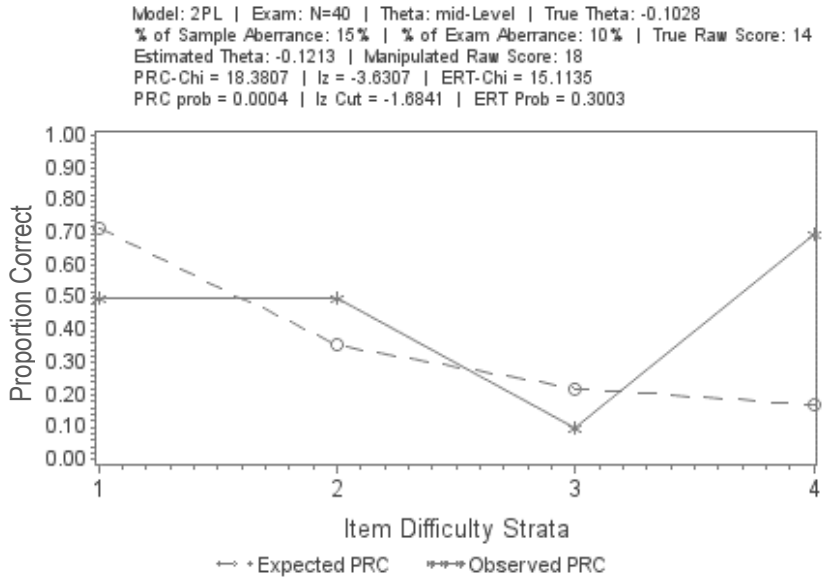

Figure A78. PRC for cheating condition: 2PL model x long form x mid ability x 15% sample aberrance x 10% exam aberrance.

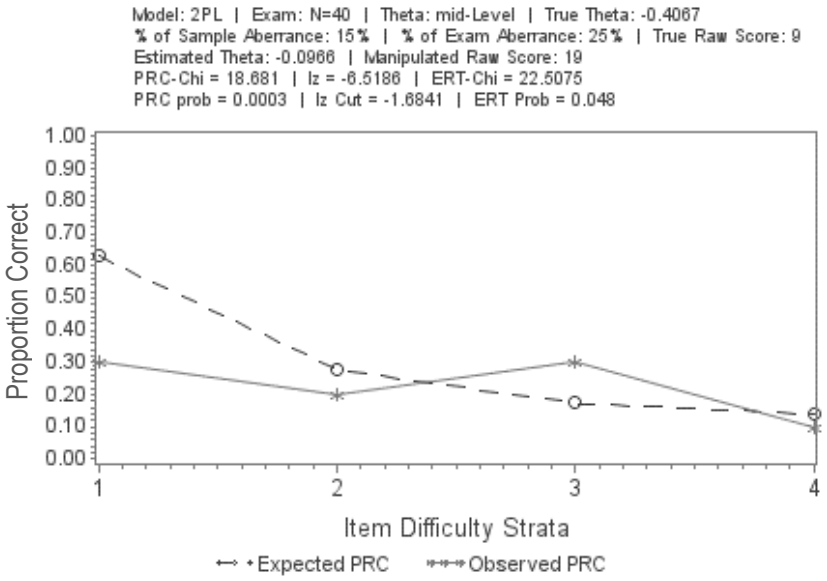

Figure A79. Baseline PRC for condition: 2PL model x short form x mid ability.

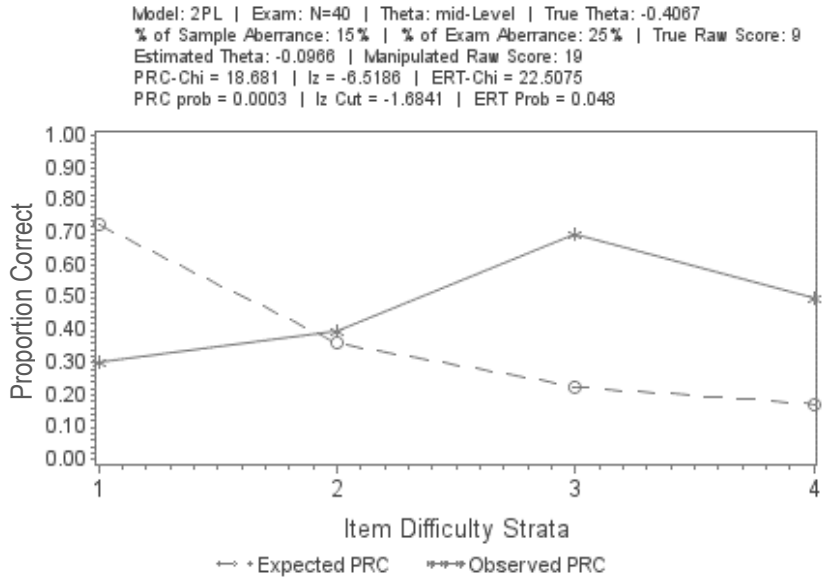

Figure A80. PRC for cheating condition: 2PL model x short form x mid ability x 15% sample aberrance x 25% exam aberrance.

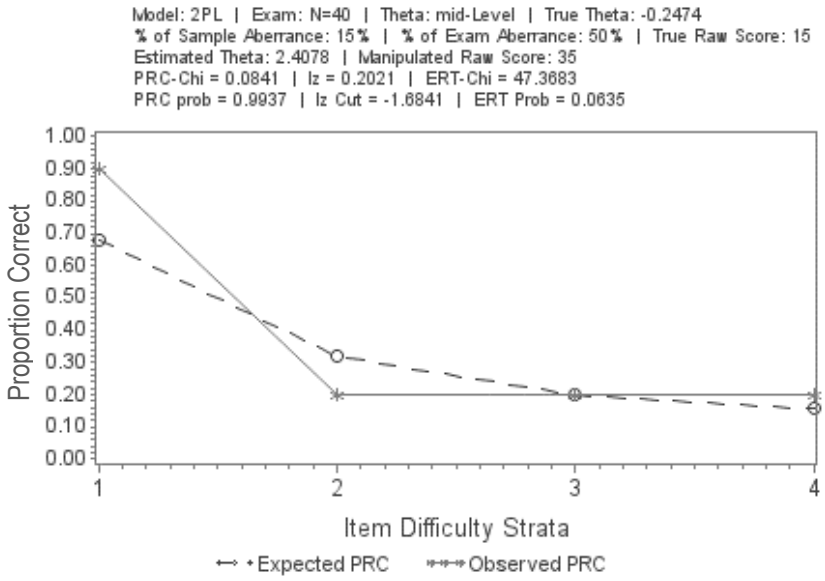

Figure A81. Baseline PRC for condition: 2PL model x short form x mid ability.

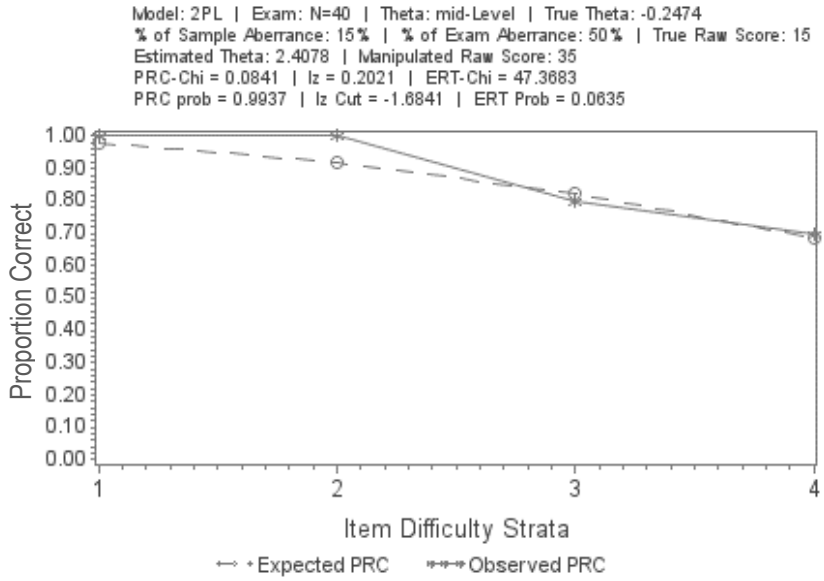

Figure A82. PRC for cheating condition: 2PL model x short form x mid ability x 15% sample aberrance x 50% exam aberrance.

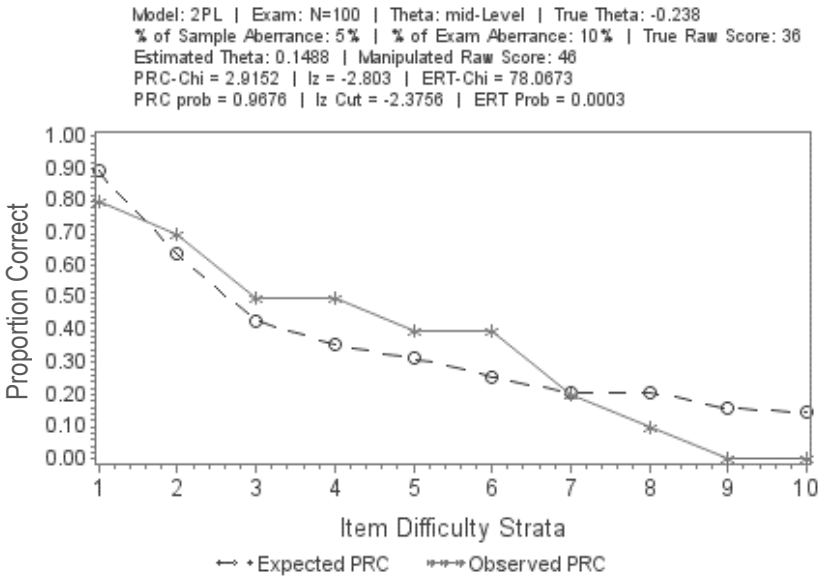

Figure A83. Baseline PRC for condition: 2PL model x long form x mid ability.

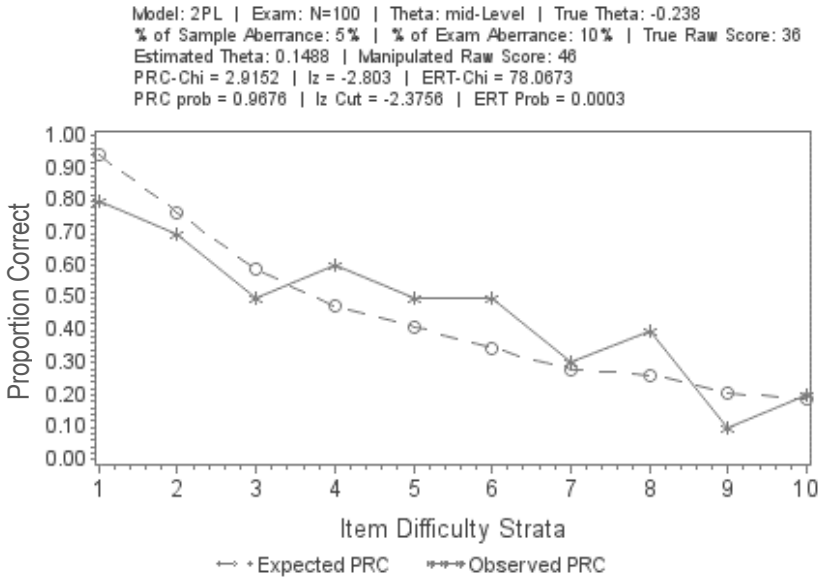

Figure A84. PRC for cheating condition: 2PL model x long form x mid ability x 5% sample aberrance x 10% exam aberrance.

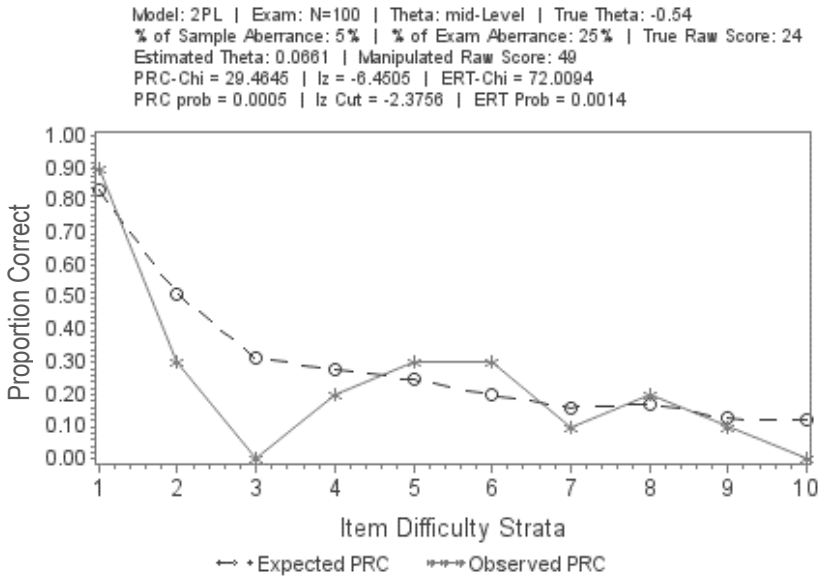

Figure A85. Baseline PRC for condition: 2PL model x long form x mid ability.

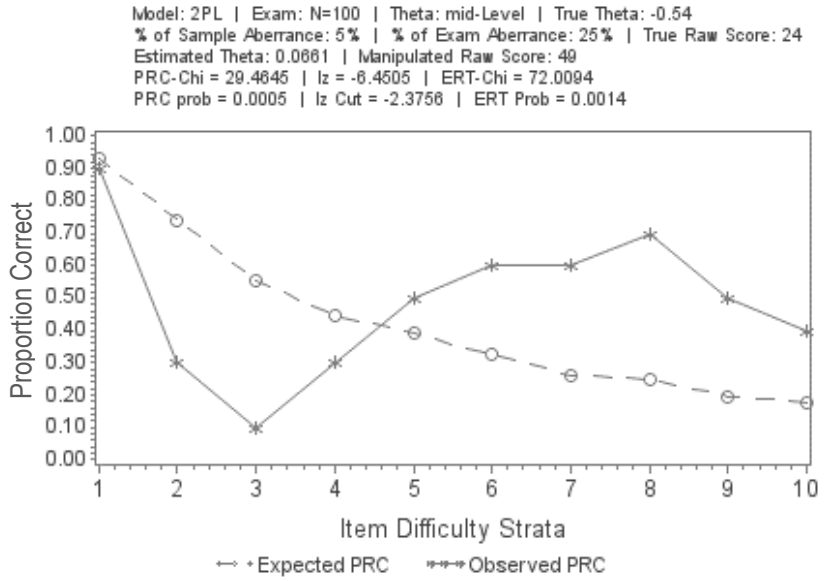

Figure A86. PRC for cheating condition: 2PL model x long form x mid ability x 5% sample aberrance x 25% exam aberrance.

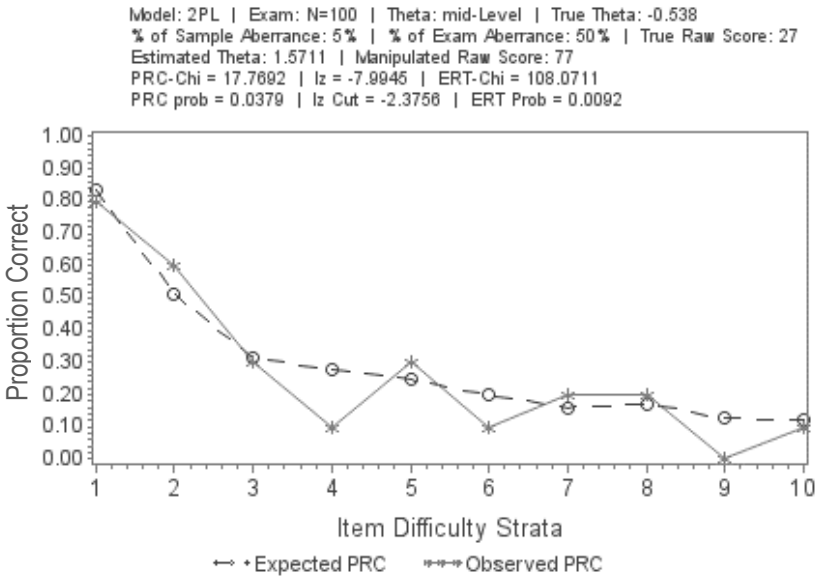

Figure A87. Baseline PRC for condition: 2PL model x long form x mid ability.

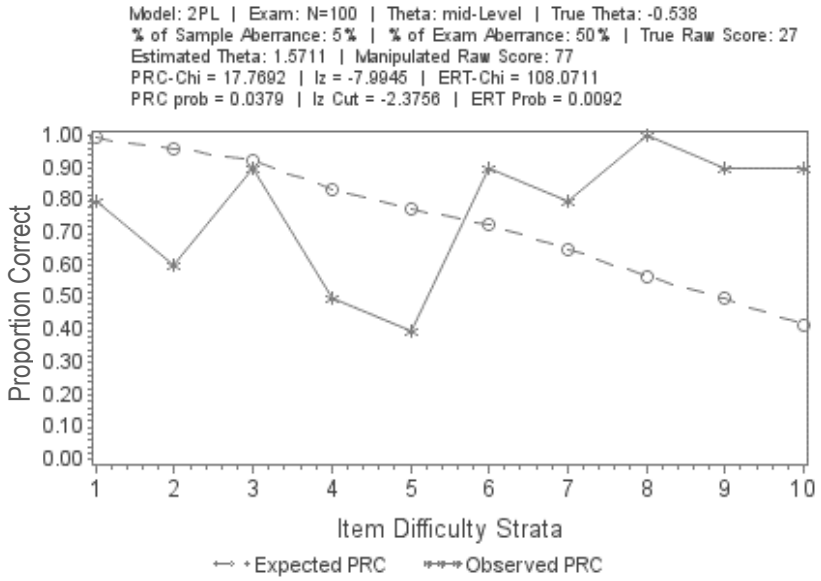

Figure A88. PRC for cheating condition: 2PL model x long form x mid ability x 5% sample aberrance x 50% exam aberrance.

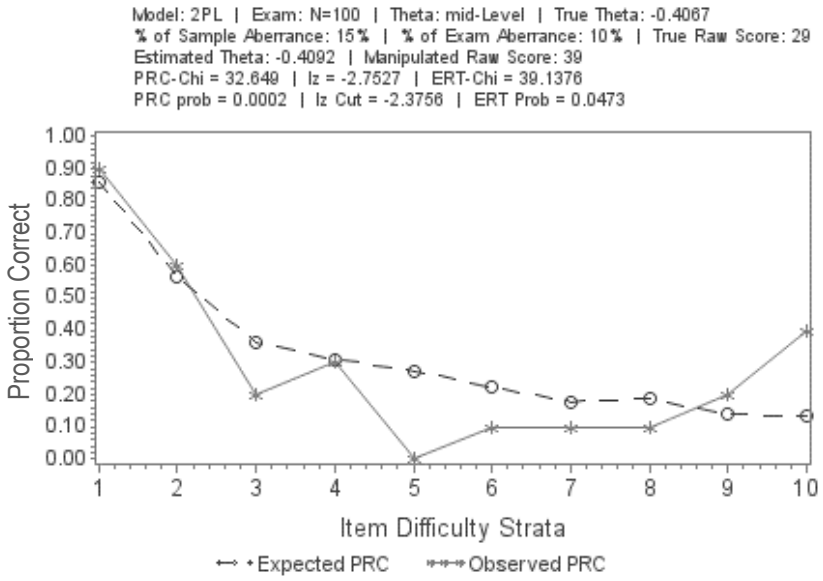

Figure A89. Baseline PRC for condition: 2PL model x long form x mid ability.

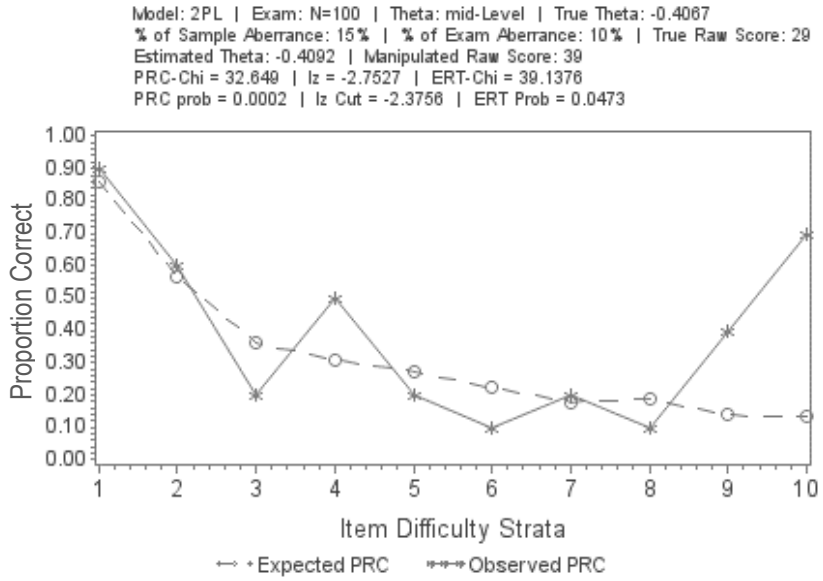

Figure A90. PRC for cheating condition: 2PL model x long form x mid ability x 15% sample aberrance x 10% exam aberrance.

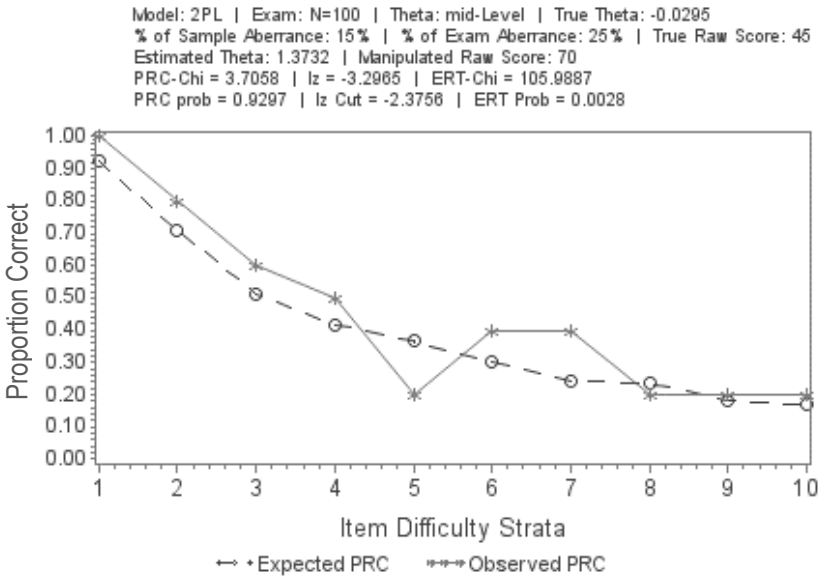

Figure A91. Baseline PRC for condition: 2PL model x long form x mid ability.

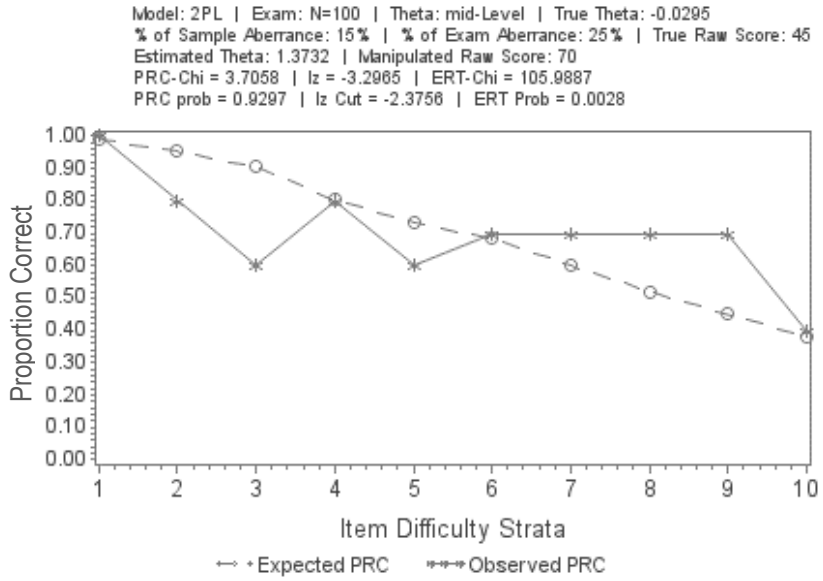

Figure A92. PRC for cheating condition: 2PL model x long form x mid ability x 15% sample aberrance x 25% exam aberrance.

APPENDIX B: Summaries of ANOVA of Type I Error Rates

Table B1. *Summary of ANOVA Results for Effect of Factors on l_z*

Effect	*df*	SS	*p*	η^2	*f*
IRT Model (M)	1	13.470	**0.000**	0.01	0.10
Exam Length (EL)	1	183.801	**0.000**	0.15	0.41
Exam Aberrance (EA)	2	236.652	**0.000**	0.19	0.48
Sample Aberrance (SA)	1	0.001	0.518	0.00	0.00
Theta (T)	1	86.564	**0.000**	0.07	0.27
M * EL	1	6.807	**0.000**	0.01	0.07
M * EA	2	2.874	**0.000**	0.00	0.05
M * SA	1	0.001	0.490	0.00	0.00
M * T	1	0.002	0.206	0.00	0.00
EL * EA	2	11.835	**0.000**	0.01	0.10
EL * SA	1	0.001	0.477	0.00	0.00
EL * T	1	3.864	**0.000**	0.00	0.06
EA * SA	2	0.001	0.790	0.00	0.00
EA * T	2	22.348	**0.000**	0.02	0.13
SA * T	1	0.001	0.365	0.00	0.00
M * EL * EA	2	4.366	**0.000**	0.00	0.06
M * EL * SA	1	0.000	0.924	0.00	0.00
M * EL * T	1	0.215	**0.000**	0.00	0.01
M * EA * SA	2	0.001	0.728	0.00	0.00
M * EA * T	2	1.058	**0.000**	0.00	0.03
M * SA * T	1	0.000	0.792	0.00	0.00
EL * EA * SA	2	0.001	0.686	0.00	0.00
EL * EA * T	2	22.744	**0.000**	0.02	0.14
EL * SA * T	1	0.002	0.190	0.00	0.00
EA * SA * T	2	0.000	0.870	0.00	0.00
M * EL * EA * SA	2	0.001	0.696	0.00	0.00
M * EL * EA * T	2	2.582	**0.000**	0.00	0.05
M * EL * SA * T	1	0.000	0.693	0.00	0.00
M * EA * SA * T	2	0.001	0.706	0.00	0.00
EL * EA * SA * T	2	0.006	0.118	0.00	0.00
M * EL * EA * SA * T	2	0.003	0.338	0.00	0.00
Error	23,952	34.400			
Total	24,000	1,256.407			

Note: Dependent Variable: Type I error α = .05.

Table B2. *Summary of ANOVA Results for Effect of Factors on ERT*

Effect	*df*	SS	*p*	η^2	*f*
IRT Model (M)	1	0.151	**0.000**	0.00	0.01
Exam Length (EL)	1	141.781	**0.000**	0.02	0.16
Exam Aberrance (EA)	2	1150.048	**0.000**	0.20	0.50
Sample Aberrance (SA)	1	63.807	**0.000**	0.01	0.11
Theta (T)	1	36.283	**0.000**	0.01	0.08
M * EL	1	0.389	**0.000**	0.00	0.01
M * EA	2	0.441	**0.000**	0.00	0.01
M * SA	1	0.010	0.028	0.00	0.00
M * T	1	1.974	**0.000**	0.00	0.02
EL * EA	2	18.484	**0.000**	0.00	0.06
EL * SA	1	0.307	**0.000**	0.00	0.01
EL * T	1	0.447	**0.000**	0.00	0.01
EA * SA	2	29.043	**0.000**	0.01	0.07
EA * T	2	78.446	**0.000**	0.01	0.12
SA * T	1	3.438	**0.000**	0.00	0.02
M * EL * EA	2	0.026	0.002	0.00	0.00
M * EL * SA	1	0.001	0.618	0.00	0.00
M * EL * T	1	0.039	**0.000**	0.00	0.00
M * EA * SA	2	0.045	**0.000**	0.00	0.00
M * EA * T	2	0.208	**0.000**	0.00	0.01
M * SA * T	1	0.000	0.675	0.00	0.00
EL * EA * SA	2	1.094	**0.000**	0.00	0.01
EL * EA * T	2	10.727	**0.000**	0.00	0.04
EL * SA * T	1	0.002	0.335	0.00	0.00
EA * SA * T	2	0.053	**0.000**	0.00	0.00
M * EL * EA * SA	2	0.004	0.437	0.00	0.00
M * EL * EA * T	2	0.141	**0.000**	0.00	0.00
M * EL * SA * T	1	0.019	0.003	0.00	0.00
M * EA * SA * T	2	0.004	0.359	0.00	0.00
EL * EA * SA * T	2	0.335	**0.000**	0.00	0.01
M * EL * EA * SA * T	2	0.005	0.303	0.00	0.00
Error	23,952	50.832			
Total	24,000	5,779.182			

Note: Dependent Variable: Type I error α = .05.

Table B3. *Summary of ANOVA Results for Effect of Factors on PRC*

Effect	*df*	SS	*p*	η^2	*f*
IRT Model (M)	1	0.100	**0.000**	0.00	0.00
Exam Length (EL)	1	124.278	**0.000**	0.01	0.10
Exam Aberrance (EA)	2	549.378	**0.000**	0.05	0.22
Sample Aberrance (SA)	1	0.000	0.859	0.00	0.00
Theta (T)	1	672.507	**0.000**	0.06	0.24
M * EL	1	0.084	**0.000**	0.00	0.00
M * EA	2	91.031	**0.000**	0.01	0.09
M * SA	1	0.000	0.915	0.00	0.00
M * T	1	3.902	**0.000**	0.00	0.02
EL * EA	2	57.799	**0.000**	0.00	0.07
EL * SA	1	0.002	0.381	0.00	0.00
EL * T	1	41.545	**0.000**	0.00	0.06
EA * SA	2	0.003	0.455	0.00	0.00
EA * T	2	54.489	**0.000**	0.00	0.07
SA * T	1	0.001	0.410	0.00	0.00
M * EL * EA	2	7.590	**0.000**	0.00	0.03
M * EL * SA	1	0.001	0.449	0.00	0.00
M * EL * T	1	0.730	**0.000**	0.00	0.01
M * EA * SA	2	0.001	0.824	0.00	0.00
M * EA * T	2	19.021	**0.000**	0.00	0.04
M * SA * T	1	0.002	0.343	0.00	0.00
EL * EA * SA	2	0.000	0.968	0.00	0.00
EL * EA * T	2	5.944	**0.000**	0.00	0.02
EL * SA * T	1	0.000	0.889	0.00	0.00
EA * SA * T	2	0.000	0.930	0.00	0.00
M * EL * EA * SA	2	0.001	0.840	0.00	0.00
M * EL * EA * T	2	3.837	**0.000**	0.00	0.02
M * EL * SA * T	1	0.001	0.438	0.00	0.00
M * EA * SA * T	2	0.000	0.907	0.00	0.00
EL * EA * SA * T	2	0.001	0.839	0.00	0.00
M * EL * EA * SA * T	2	0.001	0.872	0.00	0.00
Error	23,952	49.325			
Total	24,000	12,133.275			

Note: Dependent Variable = Type I error α = .05

Table B4. *Summary of ANOVA Results for Effect of Factors on l_z + ERT*

Effect	*df*	SS	*p*	η^2	*f*
IRT Model (M)	1	3.333	**0.000**	0.01	0.09
Exam Length (EL)	1	59.847	**0.000**	0.15	0.41
Exam Aberrance (EA)	2	117.309	**0.000**	0.29	0.63
Sample Aberrance (SA)	1	1.308	**0.000**	0.00	0.06
Theta (T)	1	11.676	**0.000**	0.03	0.17
M * EL	1	2.037	**0.000**	0.00	0.07
M * EA	2	0.963	**0.000**	0.00	0.05
M * SA	1	0.068	**0.000**	0.00	0.01
M * T	1	0.005	0.011	0.00	0.00
EL * EA	2	23.928	**0.000**	0.06	0.25
EL * SA	1	0.594	**0.000**	0.00	0.04
EL * T	1	0.732	**0.000**	0.00	0.04
EA * SA	2	0.200	**0.000**	0.00	0.02
EA * T	2	3.274	**0.000**	0.01	0.09
SA * T	1	0.458	**0.000**	0.00	0.03
M * EL * EA	2	0.934	**0.000**	0.00	0.05
M * EL * SA	1	0.047	**0.000**	0.00	0.01
M * EL * T	1	0.071	**0.000**	0.00	0.01
M * EA * SA	2	0.040	**0.000**	0.00	0.01
M * EA * T	2	0.355	**0.000**	0.00	0.03
M * SA * T	1	0.002	0.118	0.00	0.00
EL * EA * SA	2	0.365	**0.000**	0.00	0.03
EL * EA * T	2	6.125	**0.000**	0.01	0.12
EL * SA * T	1	0.173	**0.000**	0.00	0.02
EA * SA * T	2	0.027	**0.000**	0.00	0.01
M * EL * EA * SA	2	0.049	**0.000**	0.00	0.01
M * EL * EA * T	2	1.105	**0.000**	0.00	0.05
M * EL * SA * T	1	0.000	0.885	0.00	0.00
M * EA * SA * T	2	0.003	0.202	0.00	0.00
EL * EA * SA * T	2	0.083	**0.000**	0.00	0.01
M * EL * EA * SA * T	2	0.015	**0.000**	0.00	0.01
Error	23,952	19.037			
Total	24,000	408.622			

Note: Dependent Variable = Type I error α = .05

Table B5. *Summary of ANOVA Results for Effect of Factors on l_z + PRC*

Effect	*df*	SS	*p*	η^2	*f*
IRT Model (M)	1	8.998	**0.000**	0.01	0.09
Exam Length (EL)	1	173.993	**0.000**	0.16	0.43
Exam Aberrance (EA)	2	180.158	**0.000**	0.16	0.44
Sample Aberrance (SA)	1	0.000	0.602	0.00	0.00
Theta (T)	1	91.496	**0.000**	0.08	0.30
M * EL	1	6.353	**0.000**	0.01	0.08
M * EA	2	5.902	**0.000**	0.01	0.07
M * SA	1	0.001	0.313	0.00	0.00
M * T	1	0.042	**0.000**	0.00	0.01
EL * EA	2	10.488	**0.000**	0.01	0.10
EL * SA	1	0.001	0.328	0.00	0.00
EL * T	1	6.377	**0.000**	0.01	0.08
EA * SA	2	0.000	0.893	0.00	0.00
EA * T	2	25.414	**0.000**	0.02	0.15
SA * T	1	0.002	0.269	0.00	0.00
M * EL * EA	2	4.674	**0.000**	0.00	0.06
M * EL * SA	1	0.000	0.949	0.00	0.00
M * EL * T	1	0.017	**0.000**	0.00	0.00
M * EA * SA	2	0.001	0.657	0.00	0.00
M * EA * T	2	0.689	**0.000**	0.00	0.02
M * SA * T	1	0.000	0.807	0.00	0.00
EL * EA * SA	2	0.002	0.416	0.00	0.00
EL * EA * T	2	16.274	**0.000**	0.01	0.12
EL * SA * T	1	0.001	0.337	0.00	0.00
EA * SA * T	2	0.001	0.798	0.00	0.00
M * EL * EA * SA	2	0.001	0.678	0.00	0.00
M * EL * EA * T	2	1.376	**0.000**	0.00	0.04
M * EL * SA * T	1	0.000	0.650	0.00	0.00
M * EA * SA * T	2	0.001	0.668	0.00	0.00
EL * EA * SA * T	2	0.003	0.327	0.00	0.00
M * EL * EA * SA * T	2	0.003	0.300	0.00	0.00
Error	23,952	33.694			
Total	24,000	1,117.635			

Note: Dependent Variable = Type I error α = .05

Table B6. *Summary of ANOVA Results for Effect of Factors on ERT + PRC*

Effect	*df*	SS	*p*	η^2	*f*
IRT Model (M)	1	1.503	**0.000**	0.00	0.03
Exam Length (EL)	1	188.979	**0.000**	0.08	0.29
Exam Aberrance (EA)	2	382.715	**0.000**	0.16	0.44
Sample Aberrance (SA)	1	42.178	**0.000**	0.02	0.13
Theta (T)	1	51.781	**0.000**	0.02	0.15
M * EL	1	0.174	**0.000**	0.00	0.01
M * EA	2	30.830	**0.000**	0.01	0.11
M * SA	1	0.097	**0.000**	0.00	0.01
M * T	1	0.340	**0.000**	0.00	0.01
EL * EA	2	22.935	**0.000**	0.01	0.10
EL * SA	1	0.396	**0.000**	0.00	0.01
EL * T	1	0.366	**0.000**	0.00	0.01
EA * SA	2	26.373	**0.000**	0.01	0.11
EA * T	2	6.846	**0.000**	0.00	0.05
SA * T	1	5.485	**0.000**	0.00	0.05
M * EL * EA	2	1.310	**0.000**	0.00	0.02
M * EL * SA	1	0.006	0.074	0.00	0.00
M * EL * T	1	0.004	0.128	0.00	0.00
M * EA * SA	2	0.102	**0.000**	0.00	0.01
M * EA * T	2	4.402	**0.000**	0.00	0.04
M * SA * T	1	0.063	**0.000**	0.00	0.01
EL * EA * SA	2	0.586	**0.000**	0.00	0.02
EL * EA * T	2	3.178	**0.000**	0.00	0.04
EL * SA * T	1	0.002	0.340	0.00	0.00
EA * SA * T	2	0.537	**0.000**	0.00	0.01
M * EL * EA * SA	2	0.039	**0.000**	0.00	0.00
M * EL * EA * T	2	0.079	**0.000**	0.00	0.01
M * EL * SA * T	1	0.004	0.166	0.00	0.00
M * EA * SA * T	2	0.150	**0.000**	0.00	0.01
EL * EA * SA * T	2	0.387	**0.000**	0.00	0.01
M * EL * EA * SA * T	2	0.008	0.132	0.00	0.00
Error	23,952	45.377			
Total	24,000	2,394.852			

Note: Dependent Variable = Type I error α = .05

Table B7. *Summary of ANOVA Results for Effect of Factors on l_z + ERT + PRC*

Effect	*df*	SS	*p*	η^2	*f*
IRT Model (M)	1	1.731	**0.000**	0.01	0.07
Exam Length (EL)	1	54.313	**0.000**	0.16	0.43
Exam Aberrance (EA)	2	88.464	**0.000**	0.26	0.59
Sample Aberrance (SA)	1	1.252	**0.000**	0.00	0.06
Theta (T)	1	13.523	**0.000**	0.04	0.20
M * EL	1	1.700	**0.000**	0.00	0.07
M * EA	2	1.170	**0.000**	0.00	0.06
M * SA	1	0.069	**0.000**	0.00	0.01
M * T	1	0.044	**0.000**	0.00	0.01
EL * EA	2	19.910	**0.000**	0.06	0.25
EL * SA	1	0.611	**0.000**	0.00	0.04
EL * T	1	1.613	**0.000**	0.00	0.07
EA * SA	2	0.207	**0.000**	0.00	0.02
EA * T	2	4.310	**0.000**	0.01	0.11
SA * T	1	0.446	**0.000**	0.00	0.04
M * EL * EA	2	1.035	**0.000**	0.00	0.05
M * EL * SA	1	0.048	**0.000**	0.00	0.01
M * EL * T	1	0.000	0.549	0.00	0.00
M * EA * SA	2	0.040	**0.000**	0.00	0.01
M * EA * T	2	0.187	**0.000**	0.00	0.02
M * SA * T	1	0.002	0.104	0.00	0.00
EL * EA * SA	2	0.351	**0.000**	0.00	0.03
EL * EA * T	2	3.967	**0.000**	0.01	0.11
EL * SA * T	1	0.169	**0.000**	0.00	0.02
EA * SA * T	2	0.027	**0.000**	0.00	0.01
M * EL * EA * SA	2	0.048	**0.000**	0.00	0.01
M * EL * EA * T	2	0.521	**0.000**	0.00	0.04
M * EL * SA * T	1	0.000	0.868	0.00	0.00
M * EA * SA * T	2	0.002	0.204	0.00	0.00
EL * EA * SA * T	2	0.086	**0.000**	0.00	0.02
M * EL * EA * SA * T	2	0.015	**0.000**	0.00	0.01
Error	23,952	18.292			
Total	24,000	343.576			

Note: Dependent Variable = Type I error α = .05

APPENDIX C: Mean Sensitivity and Specificity Values by Study Condition

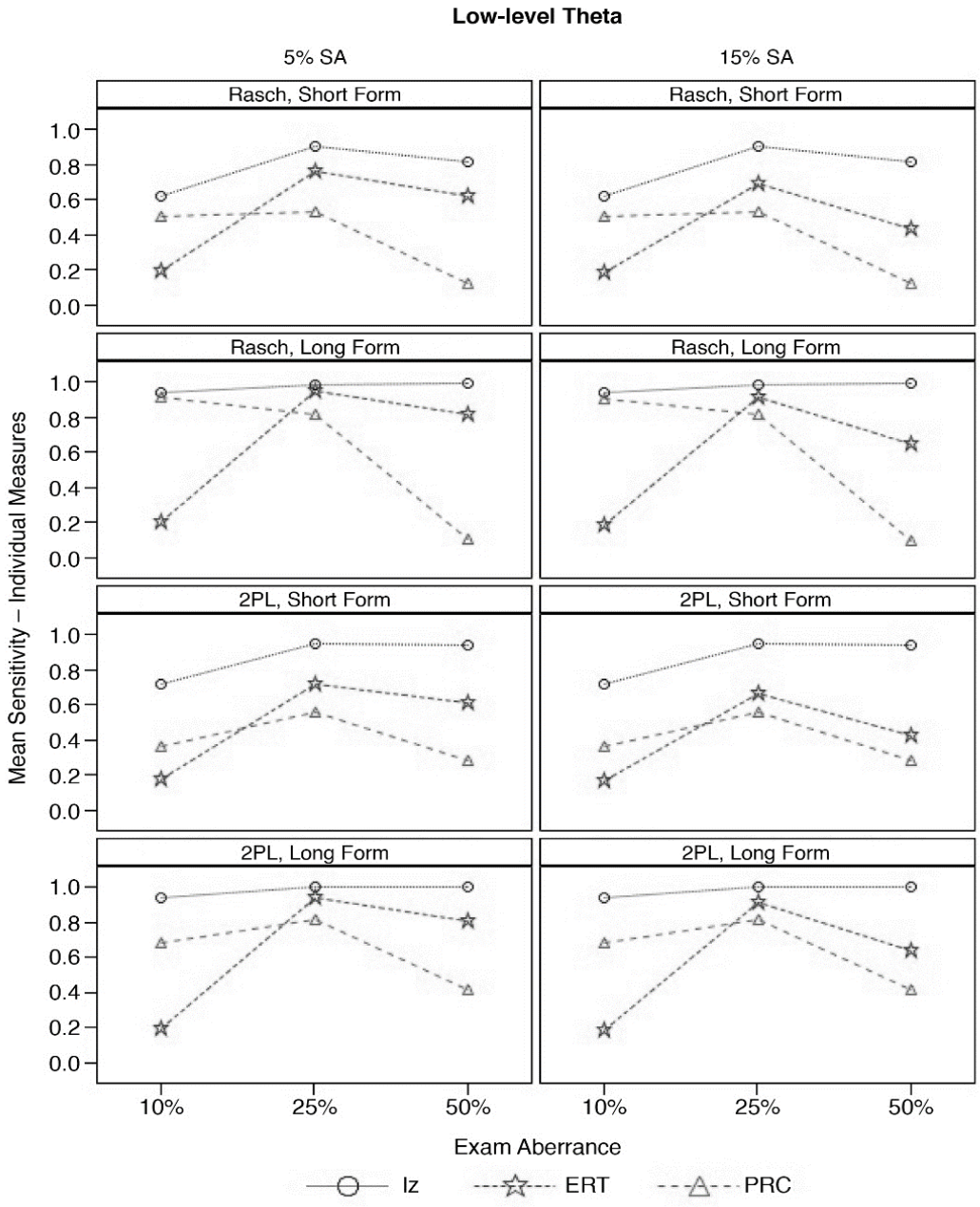

Figure C1. Mean sensitivity values for individual person-fit measures under low-level theta condition across study factors.

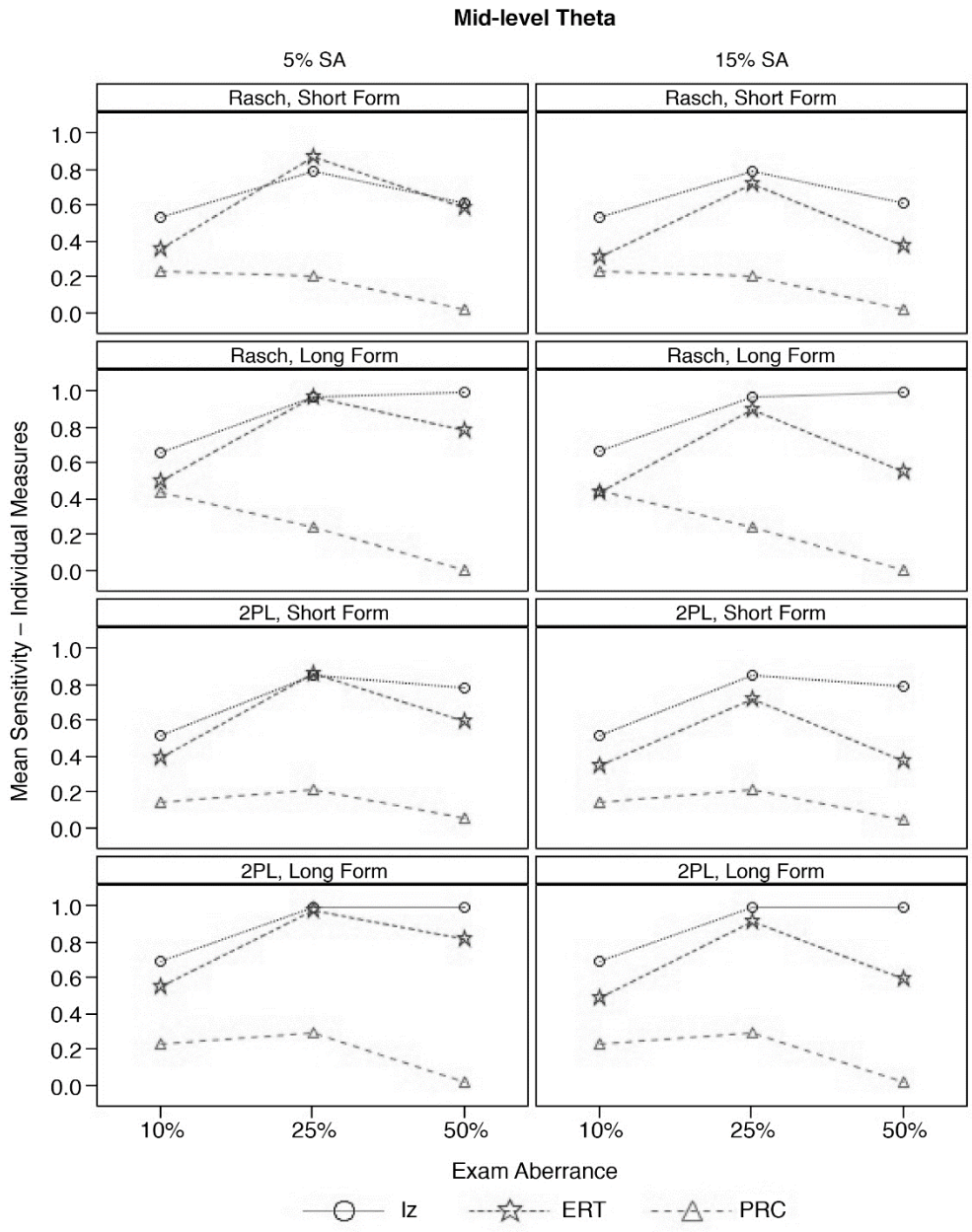

Figure C2. Mean sensitivity values for individual person-fit measures under mid-level theta condition across study factors.

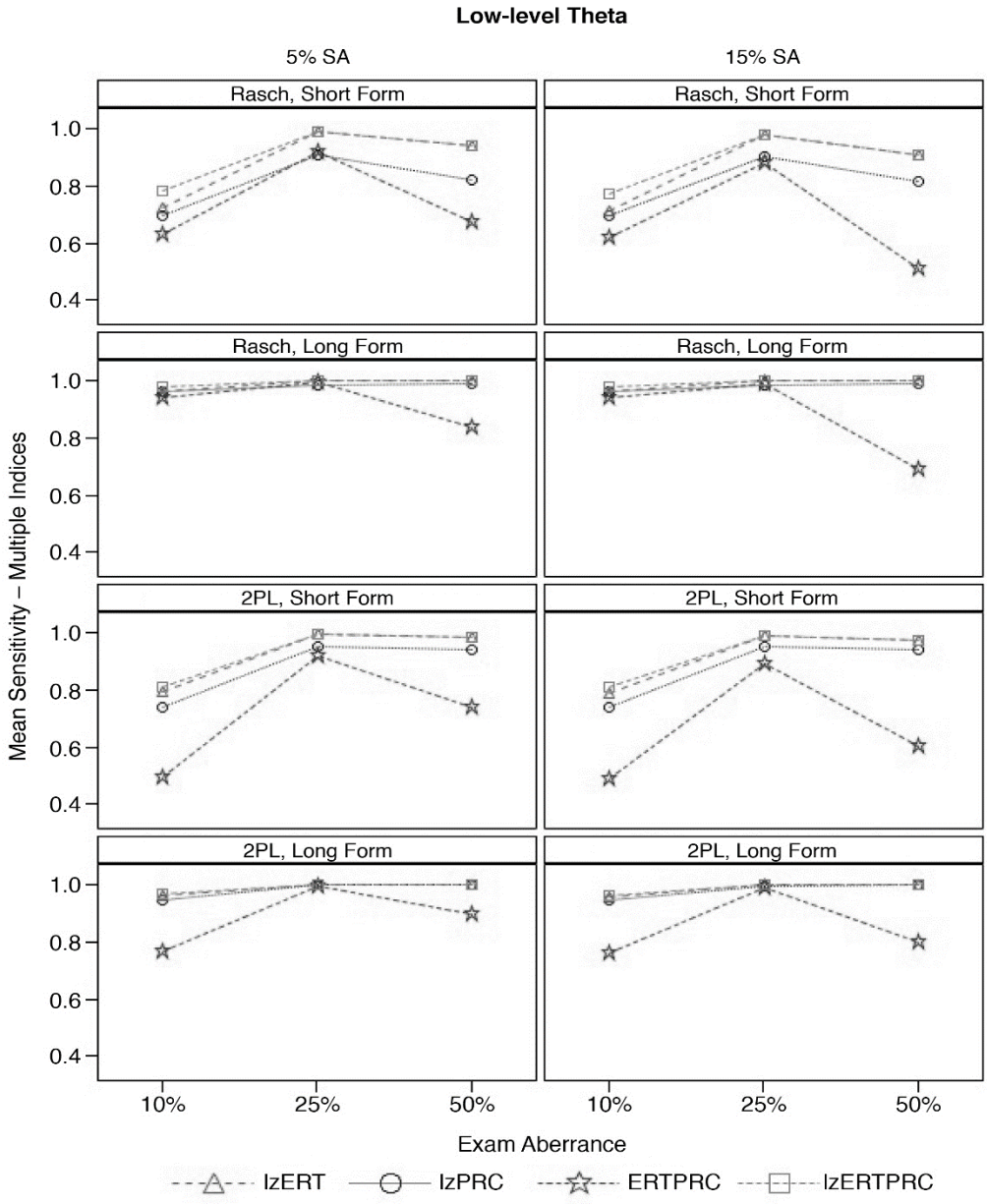

Figure C3. Mean sensitivity values for combined person-fit measures under low-level theta condition across study factors.

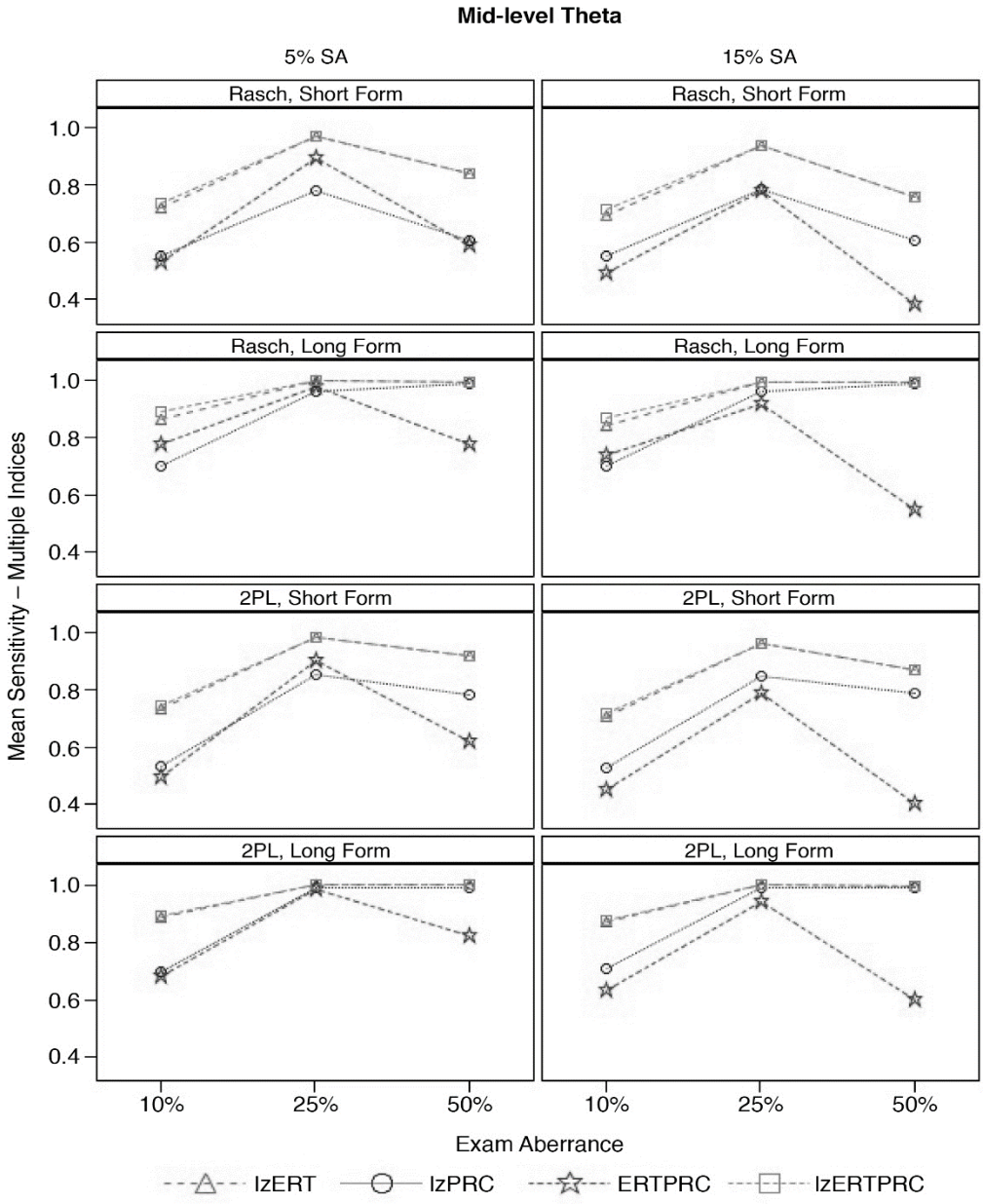

Figure C4. Mean sensitivity values for combined person-fit measures under mid-level theta condition across study factors.

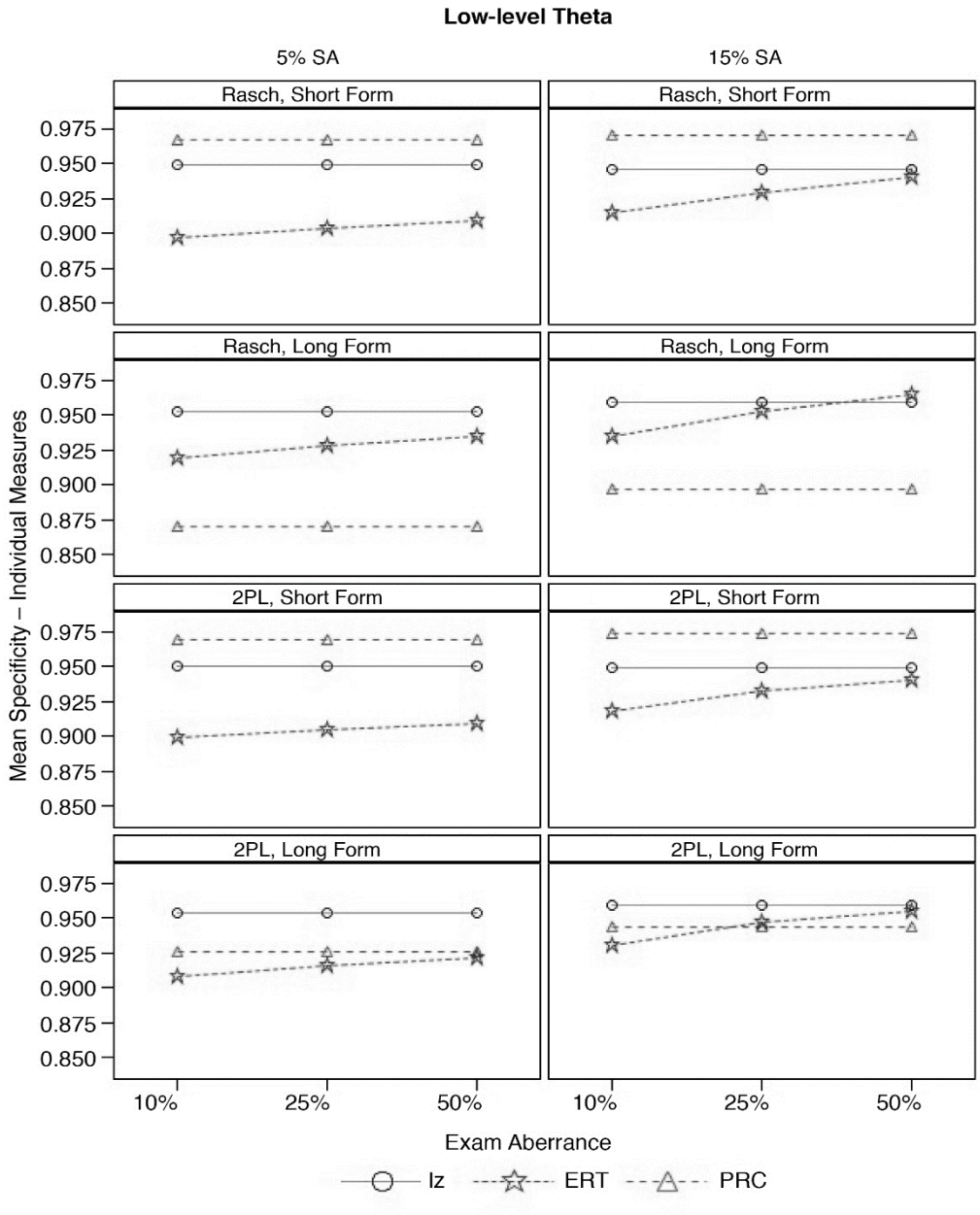

Figure C5. Mean specificity values for individual person-fit measures under low-level theta condition across study factors.

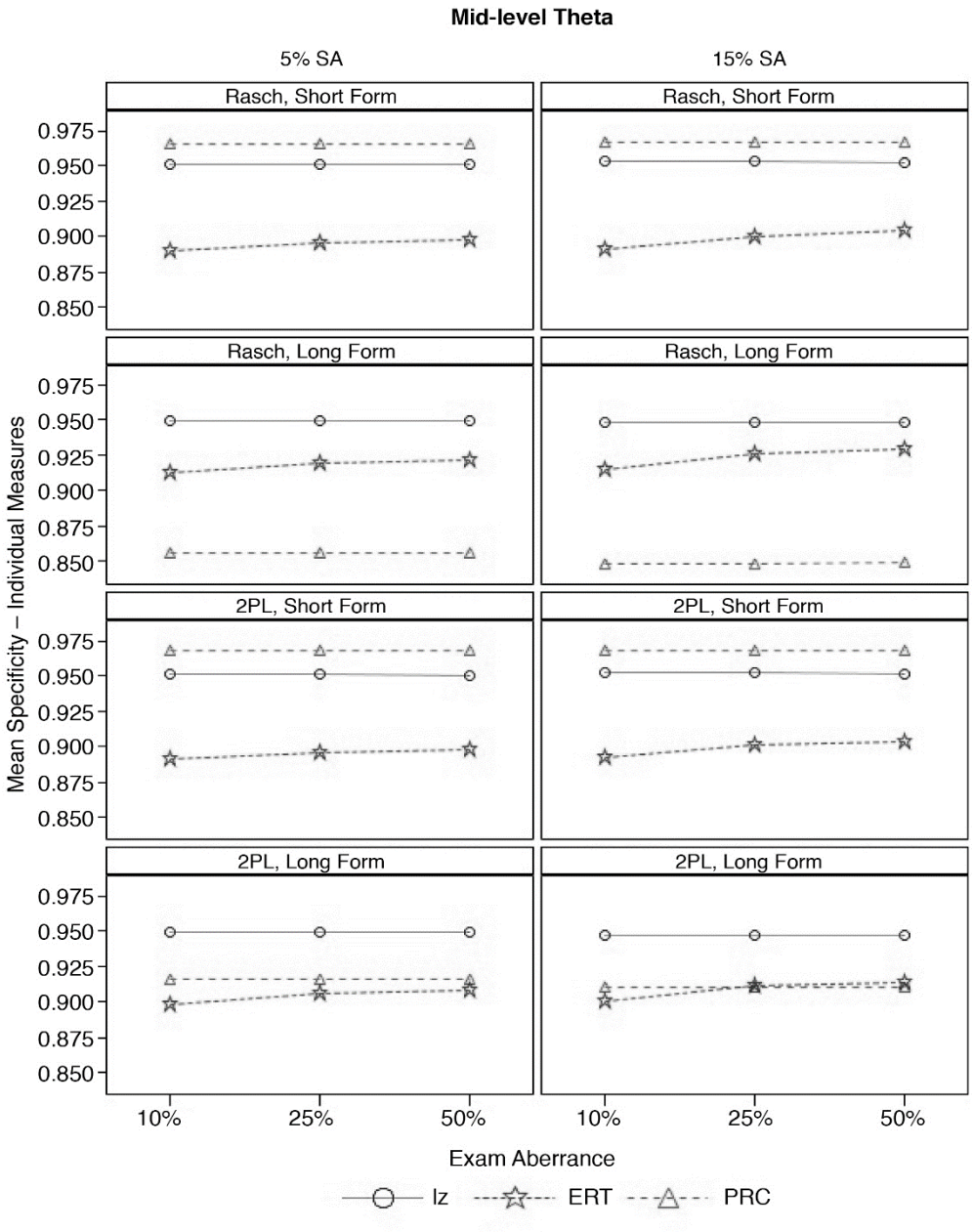

Figure C6. Mean specificity values for individual person-fit measures under mid-level theta condition across study factors.

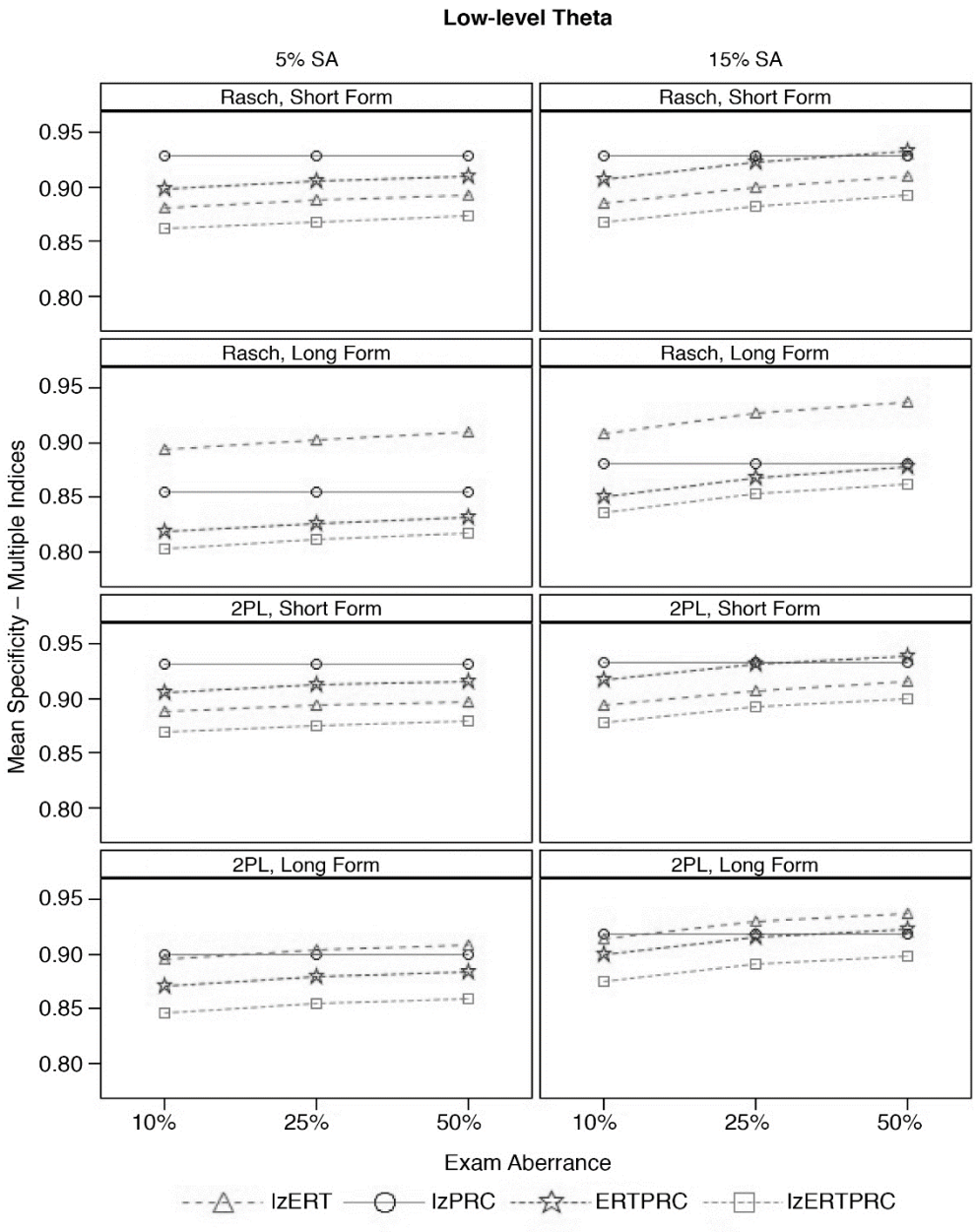

Figure C7. Mean specificity values for combined person-fit measures under low-level theta condition across study factors.

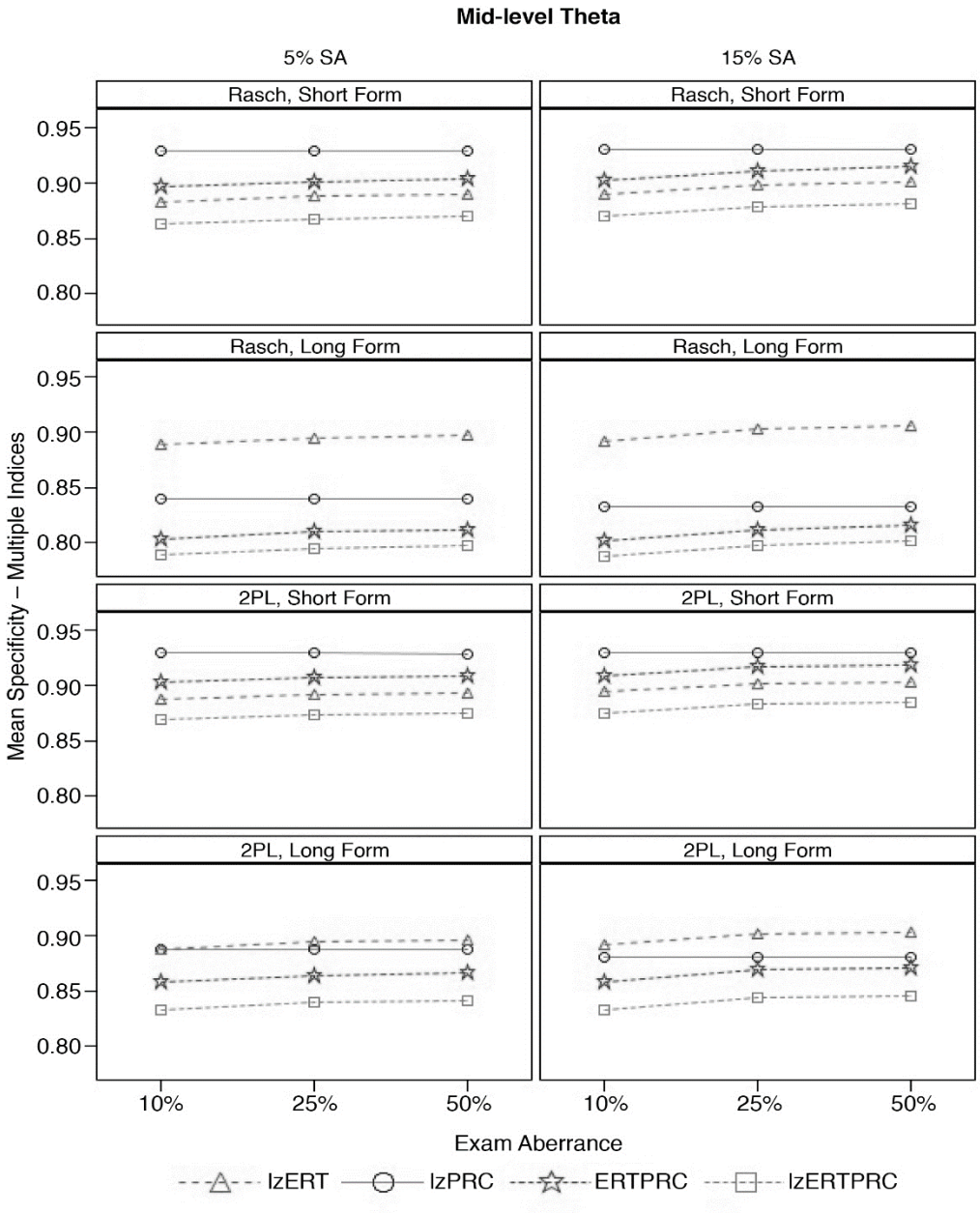

Figure C8. Mean specificity values for combined person-fit measures under mid-level theta condition across study factors.